Uncovering the Structure of Public Procurement Transactions

Mircea Popa

Forthcoming in *Business and Politics*

Close ties between government authorities and private firms are often the object of suspicion, but a systematic understanding of when they arise is still missing. This article uses machine learning tools to analyze a large dataset of public contracts from across Europe, in order to identify the conditions under which close connections, defined both in terms of repeated interaction, as well as geographical dispersion, appear. Previous theoretical results suggest that close ties should emerge as an enforcement mechanism in settings characterized by weak outside enforcement, such as those involving corruption. Results from random forest models show support for this hypothesis, along with identifying other structural determinants of the outcome. The most striking finding is that even after accounting for numerous potential confounders, major differences in terms of average diversity levels between countries persist, and these differences map onto an indicator of governance quality and corruption, but not at all on income per capita. These findings point to the centrality of the structure of interactions between private and public actors for understanding governance outcomes.

1. Introduction

Observation of market interactions between public authorities and private firms reveals substantial variation in their structure. In some contexts authorities acquire goods and services from a variety of firms, and firms similarly interact with a variety of government institutions. In others, narrower ties, characterized by repeated, undiversified interactions, dominate. The purpose of the following analysis is understanding how these differences emerge, and what they mean for theoretical

accounts of the relation between government and economic actors. In particular, the analysis asks whether the differences can mostly be explained by the technical nature of the market and the institutional framework immediately surrounding it, or whether they also point to more fundamental strategic considerations from the two counterparties. An important theoretical tradition in new institutional economics certainly points towards the second approach. According to this view, the fundamental nature of procurement interactions is a relational one, in which a particularly complex principal-agent relation exists. In such a setting, matches may form in ways determined by mutual incentives to economize on transaction costs. In particular, environments characterized by weak outside enforcement of agreements should, on average, favor the development of undiversified ties, in which familiarity allows the parties to bring predictability to their interaction. Of particular concern for the public procurement setting is the situation in which the weak outside enforcement is due to the corrupt, or otherwise socially undesirable, nature of the transaction. In such settings, repeated or otherwise close interactions between public and private agents could perpetuate these undesirable outcomes. The analysis in the following therefore seeks to identify the conditions under which close ties between public authorities and private firms develop, and to evaluate whether there is indeed a connection between such close ties and undesirable outcomes, whether measured at the aggregate country level, or at the level of the individual transaction.

Open-government data on public procurement in the European context will allow an empirical analysis of patterns relevant to this question, the extent of which, to our knowledge, is novel to the literature. The statistical analysis will make use of a dataset on 3.3 million public contract awards from 33 European Economic Area members and associate countries, between 2009 and 2015. The connection between the diversity outcomes and their predictors will be

estimated through random forest models (Breiman 2001), which have been developed in the statistical learning or ”machine learning” literature, and which have significant advantages when the objective is an accurate modelling of the outcome in problems with lots of variables and little guidance about the true functional form of the model. Their interpretation however is similar to that of other statistical models, and much of the technical detail has been relegated to the appendices. As a secondary technical consideration, significant effort has gone into forming unique identifiers for the firms and government authorities in the data, given that a fully reliable method for identifying them does not exist. Again, much of the detail is found in the appendices.

The results of the analysis point towards a strong connection between the governance environment and the structure of matches, whether in terms of repeated interactions or geographical distance. The structural determinants of the outcome, such as the nature of the product or the type of buying authority behave as expected, but still allow for significant variation to be explained by more theoretically relevant variables. The most striking finding is that, even after accounting for a wide variety of other predictors, the structure of matches still differs greatly between countries, and those differences map onto an indicator of the quality of governance: countries with cleaner and more effective government feature higher average levels of diversification of the matches between public and private actors. Moreover, the connection is not explained by different levels of economic development, which are unconnected to the outcome once governance quality is accounted for. Beyond this, less diversified ties are also predicted by contract-level indicators of undesirable outcomes, such as less competition, and to some extent less open bidding procedures. These patterns offer support for the idea that undiversified ties are integral to the functioning and survival of inefficient and corrupt systems of governance; and complement previous theoretical, qualitative, and experimental works on this topic. More

generally, the findings offer empirical support for a key claim of the new institutionalist literature, namely that repeated interaction should be expected to emerge as an enforcement mechanism in settings characterized by weak outside enforcement.

This article contributes to an emerging literature on the political economy of public contracts, and broadly complements existing findings in these works. Boas et al. (2014) show that public contracts are a key driver of corrupt exchanges between business and politicians in Brazil. By contrast, Aggarwal et al. (2012) show that electoral donations have no effect on the awarding of public contracts in the US. These contrasting findings justify a focus on examining the connection between governance quality and the nature of procurement interactions. Charron et al. (2017) show that corruption markers in contracting data (single-bid contracts, restricted procedures, and others) are connected to the career incentives of the bureaucrats awarding them, with more political control predicting more problematic outcomes. Klasnja (2017) uses markers of corruption in Romania (including discrepancies in asset disclosures, indicators of suspicious contracting procedures, and public spending data) to test for their effects on incumbency disadvantage, and finds a substantial impact. Lonsdale et al. (2016) provide a careful empirical analysis of opportunistic behavior on the part of suppliers, founded in the same transaction-cost arguments as here. Hansson (2012) similarly analyzes the opportunistic behavior of public authorities in the context of EU procurement, and the private sector response to this. Baldi et al (2016) analyze the connection between project complexity, institutional framework, and corruption, in an Italian setting. Fazekas and Koksis (2017) develop a methodology for identifying corruption in contracting from institutional markers (including awarding without a call for tenders, restrictive procedures, short time frames, and subjective evaluation criteria), which will be useful for interpreting the results in this paper. This article complements these works by focusing on a

factor which has received comparatively less attention, namely the diversity of ties between buyers and sellers, and discussing the connections between this and other key variables from the literature. Section two of the article will present the theoretical background of the analysis and its connections to existing literature; section three will present the data, together with the random forest methodology; section four will present the results, and section five will offer some conclusions.

2. Theory and connections to the literature

The political economy literature on government - firm interactions in the procurement context draws upon contract theory and new institutional economics. There is wide agreement in the literature that the procurement transaction is characterized by a complex principal-agent problem involving the buyer, the seller, and the public as a whole (Laffont and Tirole 1993, Bajari and Tadellis 2001, Spiller 2009). The first aspect of the problem is the relation between the government actor and the business. Transaction costs in this relation arise from the possibility of opportunistic behavior on the part of the firm, the government authority, or even third parties. Possible solutions to the problem include repeated interaction (Rey and Salanie 1990, Corts and Singh 2004, Corts 2011) and reputation-building (Banerjee and Duflo 2000, MacLeod 2007). This provides the first reason we expect environments with weaker outside enforcement of agreements to lead to less diverse ties, if such ties emerge as solutions to commitment problems. At this level there would be nothing necessarily corrupt about such ties, as they could be merely an adaptation to an adverse institutional environment.

It is unlikely however that the story ends here. The second aspect of the problem refers to agency from the public towards the authority-firm pair (Lambsdorff 2002, Della Porta and Vanucci 2004). As the authority and the firm are spending and receiving somebody else's money,

in settings with weak outside enforcement there are strong incentives towards collusion and mutual extraction of rents from the transaction, by, for example agreeing on an excessive price or tolerating poor quality. These rents could then be distributed between the public and private actors. An extreme form of this arises when the government actor is effectively dealing with herself, in situations in which the firm is under her control. At the other end of the continuum the collusion can take the subtle form of a cozy relationship, in which substantial inefficiency exists, but officials are spared the effort of searching for and developing new connections, and the firm derives supercompetitive profits. The key characteristic of such interactions is that they breach the public's trust, and therefore their illicit aspects are not subject to outside enforcement (Lambsdorff 2002, Lambsdorff and Teksoz 2004, Kingston 2007). A series of works have argued that interactions lacking third-party enforcement should lead to undiversified ties being formed, in which repeated play is the chief enforcement mechanism. The argument has been made on a theoretical level (Klein and Leffler 1981, Shapiro and Stiglitz 1984, Hart and Holmstrom 1987 are some foundational references), as well as tested in an experimental setting (Brown et al. 2004). In a more applied setting, Tonoyan et al. (2010), as well Jancsics and Javor (2012) argue in two studies of corruption in Eastern Europe that close social ties are a chief enforcement mechanism for illegal interactions in the region. The literature on the negative effects of social capital (Portes and Landolt 1996, Rosenbaum et al. 2013, Murray et al. 2015) similarly cautions that while close social ties between pairs of actors can facilitate cooperation between them, this does not imply the social desirability of such cooperation. Similar conclusions could be derived from the sociological literature on weak ties (Granovetter 1977), which argues that diffuse, numerous ties, between agents can lead to better economic outcomes; as well as from the distinction between particularism and universalism in characterizing the fundamental nature of corrupt interactions present in the

political science literature on the topic (Mungiu-Pippidi 2006, 2013, Rothstein 2011). This provides the second reason why we expect markers of poor governance to be connected to undiversified interactions, as the undiversified ties should emerge as a socially undesirable adaptation mechanism.

The two channels suggested above could, in principle, manifest themselves separately: we could imagine a situation in which the close ties emerge only through the first mechanism, when fully uncorrupt and efficiency-minded officials, along with law-abiding firms, engage in repeated or otherwise close interactions due to poor enforcement of agreements by the judiciary. This however is unlikely in practice. An environment in which opportunistic behavior towards the counterparty to a transaction is not well policed is very likely also an environment in which opportunistic behavior towards the public is not well policed, making the distinction moot. Going even beyond that, Lambsdorff and Teksoz (2004) make the argument that connections between public and private actors that emerge for legitimate reasons then generate the environment of trust which facilitates the development of corruption. Once the trust between parties has emerged in a setting of weak outside enforcement, the assumption that it will not be used for mutual income maximization would be hard to sustain. For all these reasons, it would be difficult to argue that the connection between poor governance and close ties is indicative of a socially "second-best" adaptation.

An aspect of the diversification of ties which has not received as much attention in the literature is their geographical distribution. When transaction costs increase with distance (as would be the case in a setting where joint, illegitimate, rent extraction is the objective of both parties and therefore impersonal, long-distance, agreements are hard to maintain), local ties will be favored by officials. Such interactions may be easier to maintain in the absence of outside

enforcement, and may arise naturally when the buyer and the seller are just two instances of the same entity. Local ties would also emerge when the motivations of political actors in favoring local companies are political but not directly extractive in nature, for example when they wish to support local employment and/or the success of local donors (see Eggers and Hainmueller 2013 for this dynamic in the case of the US). If indeed geographical diversity plays a similar role to our previous conceptualization of diversity, we would expect the predictive model for this outcome to behave similarly to the first case. Indications of this logic are present in the literature on parochial corruption (Kingston 2007), as well as on the governance of illicit transactions (Lambsdorff 2002, DellaPorta and Vanucci 2004), even if not explicitly spelled out.

The economic logic outlined above provides one motivation for studying the emergence of diversified versus undiversified ties. If the logic is valid, then undiversified ties should disproportionately emerge in countries with poorer governance, and should also be associated with contract-level markers of socially undesirable outcomes, as identified by previous literature. Undiversified ties would then be both a cause and an effect of such outcomes. They would arise when agents are intent on acting in such socially undesirable ways and the wider institutional environment does not provide a check on their intentions, and once formed they would sustain collusive behavior on the part of the buyer and the seller. While this logic is relatively simple, due to data limitations it has received limited empirical support so far. Brown et al. (2004) tackle a part of this claim in an experimental setting, and show that indeed undiversified, repeated ties, emerge naturally in transactions without third-party enforcement. Extending this result to representative observational data would therefore strengthen these conclusions and confirm that a basic proposition of the theoretical literature does indeed hold in real-world data. (As also noted by Brown et al., this is especially important as conclusions regarding cooperation under repeated

interaction are derived from models that almost always generate multiple equilibria, and there should be no a priori assumption that the cooperative one is generally chosen.) Moreover, extending the results to a geographical understanding of diversification would point towards the same logic being at work here, and towards the relevance of geographical proximity to our understanding of inefficient or corrupt interactions.

A different strand of literature relevant to our argument looks at the effects of known ties between firms and officials on firm performance. The conclusions of this literature are generally that such ties do lead to supercompetitive returns, in settings as varied as the US (Goldman et al. 2008), Brazil (Claessens et al. 2008), Pakistan (Khwaja 2005), Hungary (Fazekas and Toth 2016), and cross-nationally (Faccio and Parsley 2009, Boubakri et al. 2012). A notable exception to this conclusion is Fisman (2001), who argues that in a setting with strong rule of law, the US, such ties did not lead to excess returns. These findings further justify attention towards mechanisms that may strengthen ties between firms and public authorities, such as repeated interaction.

Testing the above propositions in observational data is not trivial because the equilibrium nature of the ties between buyers and sellers will very likely be influenced by a host of other structural and economic factors. The nature of the product being transacted is an obvious one: some markets, especially those for complex products, are simply more concentrated on either the seller side, or the buyer side, or on both (Brown et al. 2009). It may also be that various types of government authorities (such as central government ministries, local government authorities, or public utilities) behave systematically differently in these transactions, for reasons which have little to do with the logic above. Including such factors in any explanatory model is therefore warranted for meaningful conclusions to be drawn. It may indeed emerge from the analysis that most, or all, of the variation in the nature of firm - authority ties are due to such structural and

economic reasons, which, while interesting to analyze in itself, would cast doubt on the relevance of the outcome for wider questions regarding governance and efficiency. The same arguments apply to the geographical distribution of ties.

An alternative which is even farther removed from the governance and transaction costs argument is one in which undiversified ties are simply signs of efficiency: If buyers manage to identify the best suppliers and sellers similarly manage to specialize in serving the buyer for which they can do the best job, then repeated interactions between buyers and sellers would not be a sign of an environment with high transaction costs, but simply of first-best efficiency. (This would certainly be the view adopted by public officials and firms quizzed on suspiciously close ties). If this view is valid, then we would expect the opposite patterns to hold in the data, that is close ties should be associated with positive outcomes, which would cast doubt on the applicability of the transactions-cost view, at least in this European setting.

The empirical effort motivated by the arguments above is one in which contract-level measures of tie diversity (whether in terms of repeated interaction or close geographical proximity) are first studied as the outcome, and a host of competing contract-level explanatory factors are used as predictors, in addition to country fixed effects meant to model the average diversity level for each country. Separate country-level models can then be used to check whether these country-level averages of the diversity outcomes are indeed associated with indicators of governance quality and other country-level controls.

3. Data and methods

The full dataset comprises all public contracts which have been published in the Journal of the EU between January 2009 and December 2015. There are 3,307,700 contract awards, from 33 countries, including all EU member states, the members of the EEA, and two candidate countries, one of which joined the EU during the period. The reliability of the data is supported both by the legal requirement regarding publication of public contract calls and award notices worth beyond certain monetary thresholds in the Journal (arising through Council Directive 2004/18/EC, updated by Council Directive 2014/24/EU), and by the fact that it is used by the European Commission for policy analysis (PwC, London Economics, and Ecorys 2011, European Commission 2016). The most relevant thresholds, are €133,000 in 2009, rising to €135,000 in 2015 for most contracts, and €5,150,000 in 2009, rising to €5,225,000 in 2015 for infrastructure projects (European Commission 2016). These values refer to the total value of the contract, but contracts are often split into lots, also called "contract awards", which will be of lower value.

The Journal entries are legal documents, and therefore the quality of the winner and authority data recoded in them can be expected to be quite high. The forms require the "official name" of the winning company, as well as of the contracting authority to be recorded. However, the nature of the recording process, done by potentially thousands of different employees across a country, means that inconsistencies are inevitable. Moreover, some companies may have several operational divisions, and it is not clear whether the division or the larger company should be recorded in these fields, providing a further source of potential error. The "record linkage" task of merging different recordings of the same entity has received significant attention in computer science (Christen 2012). The procedure used here follows the basic steps from the literature, with the full algorithm being described in appendix 4. To test the success of the procedures, a random

sample of 100 contract awards was extracted from the full dataset and analyzed manually, with the results presented in table 1.

The first step of the linkage algorithm is a cleaning of the data to remove capitalization, punctuation, and common designations such as “Inc.” or “SA”. This step reduces the number of unique names by 33% for companies and 24% for authorities, and generates classification accuracy levels of 84% and 89% on our test sample of 100 cases, respectively. The second step is to make use of the address information provided in the forms. While sharing the same street address *and* a similar name at the same time is not a necessary condition for a match, it is arguably a sufficient one. The third step is clustering similar names together based on a measure of string distance. The procedure uses the Jaro-Winker distance (Jaro 1989, Winkler 1990), which has been shown to be the most accurate for name-matching tasks by Cohen et al (2003). The clustering algorithm is based on the logic that similar names should be grouped together, and that the more frequently encountered one is more likely to be the correct one. Therefore, for every unique name in the dataset the algorithm searches for the closest match among the more frequently encountered terms, and links the entry to the more common one if the distance is below a certain threshold.

Table 1 presents the estimated accuracy of four procedures on our sample of 100 contract awards: In each case, an entity is recorded as correctly classified if it avoids both a false positive and a false negative error. Additionally, the two joint bids in the sample are always counted as misclassified. In this and all other linkage procedures, a tradeoff between avoiding false negatives and false positives will arise. While the mild cleaning of the data generates no false positives, it will obviously miss many matches. As more aggressive joining criteria are used, the balance shifts towards more false positives. The table shows that the algorithm achieves an estimated overall accuracy of 95% for company names, for a clustering with a threshold of .05, and of 97% for

authority names, when they are linked on the address and similarity. The body of the paper presents results on this combination of parameters, and appendix 2.3 provides results on the other combinations, to show that movements along the tradeoff between false positives and negatives do not affect the basic results, beyond creating more noise in the data, which is reflected in slightly lower predictive accuracy. The errors that survive the linkage procedure are unlikely to affect the findings beyond introducing noise in the estimation process because they are based on considerations of language rather than on theoretically relevant factors, and are likely orthogonal to the patterns uncovered in the analysis. Appendix 4 provides more details on this record linkage procedure, including a description of less successful attempts.

Contract winners	"Cleaned"		"Address-merged"		"Clustered .05 distance"		"Clustered .10 distance"	
Classification accuracy	.84		.92		**.95**		.92	
Pos/neg accuracy	1	.84	1	.92	1	.95	.96	.92
Unique entries	720,080		620,518		559,683		441,805	
Contract authorities								
Classification accuracy	.89		**.97**		.95		.91	
Pos/neg accuracy	1	.89	.99	.97	.97	.95	.93	.91
Unique entries	122,380		101,859		95,738		82,294	

Note: Results based on a sample of 100 contract awards. Sampling seed 12345 in R. Joint bids (.02 of sample) are counted as misclassified in all cases. The first cell for "pos/neg accuracy" is the percentage correctly included in its cluster. The second cell is the percentage not included in the correct cluster. If an entry fails the first criterion, it also fails the second. Overall accuracy is percentage meeting both criteria.

Table 1: Estimated accuracy of record linkage procedure

Another methodological concern is that some of the variables contain missing data. This most often arises not as a result of willful misreporting, but because the quantity does not apply to that transaction. For example, contract awards for which the total price is not established beforehand will not have a price being recorded, and so on. In these cases it would be inappropriate to impute the values, so, in order to make sure they are included, missing data is always treated as a separate category for categorical variables. (This is also sometimes done by default in the EU

data). The two continuous variables are transformed into a set of indicators for the quintiles and deciles of the distribution, respectively. This allows treating the missing data as a supplementary category. Given the excellent ability of the random forest models to deal with such categorical variables, this should not generate any meaningful loss of information.

To model the connection between explanatory factors and the dependent variables, a random forest (RF) model is especially appropriate given the nature of the data. RF is a machine learning technique based on decision trees and bootstrapped aggregation of the results of multiple trees (see for example Hastie et al. 2009). A decision tree is series of bifurcations that subdivide the sample according to splits on the independent variables. The splits are performed according to the criterion of minimizing squared loss in the dependent variable, and they take place until a small number of data points are present in each terminal node of the tree. (Fifty data points in each terminal node works well for this very large sample, and there is no practical advantage in growing trees which are deeper than this). While a single decision tree can provide a good model of the data, the predictive accuracy of the model can be improved by aggregating the results of many trees (200 in our case), each estimated on a bootstrapped sample, which provides a predictive model with less variance than considering just one tree. As each bootstrap sample leaves some observations outside of the sample, a cross-validation exercise can be automatically performed, which means random forests also offer protection against over-fitting the data. Additionally, this bootstrapping process allows the estimation of standard errors for our measures.

The RF model has a number of advantages compared to traditional linear models given the nature of our data. First random forests automatically take into consideration possible nonlinearities in the data, which in problems with many variables, each with a large number of categories, would be impossible to do in a systematic way using linear regression. Our data is

especially complex, featuring a mix of continuous and categorical variables, some with hundreds of levels, which makes this problem especially salient. Second, RFs have the advantage of producing a simple measure of variable importance, which summarizes the total effect of one variable on the outcome, across all of its interactions and other nonlinearities, and allows to test for the overall significance of the variable independent of any functional form assumption. This again is useful for the problem at hand, as we are interested in the degree to which various predictors are meaningful explanatory factors of the diversity outcome independent of any linearity assumption. Third, in order to gain insight into the behavior of each variable in the model, random forests can generate a plot of its average partial effect, which is similar in nature to those obtained from traditional linear models. This allows for easy interpretation of the direction and magnitude of its marginal effect. As a robustness check, appendix 6 presents the main models estimated using linear regression, and assuming a simple additive functional form. These results are similar in substantive terms to the ones from random forest models, offering reassurance that the findings are not an artefact of the statistical tools.

The first set of estimates come from models in which the diversity dependent variable is defined in terms of repeated interactions. The full dataset contains separate entries for each transaction *i*, between firm *f* and authority *a*. The dependent variable for all transactions between *f* and *a* is therefore the total number of transactions between them recorded in the sampling period. As this relationship-level outcome does not vary among the component transactions, it creates dependence between the data points, which may affect the precision of our estimates (Adler et al. 2011, Karpievitch et al. 2009). Adler et al (2011) propose as a simple solution to this problem, in the context of random forest models, sampling one data point among those with a common

dependent variable, in this case a single transaction[1]. Models will therefore be estimated on a dataset resulting from such random sampling of one transaction per *f-a* relationship[2]. For the number of matches to capture our understanding of diversity, it needs to be conditioned on the total number of transactions that both *f,* and respectively *a* engage in in the sample, as larger authorities and larger firms may interact more frequently simply due to size. Therefore these two quantities, denoted firm award count and authority award count, are always included among the predictors. As all three variables are right-skewed, they are transformed through a natural logarithm.

A second set of results will use as a dependent variable the geographical distance between buyer and seller, instead of the number of matches, while using the same set of predictors. One transaction per *f*-a relationship is sampled here as well, with the same justification. The log distance between the cities recorded for the buyer and seller in each transaction is computed and used as a dependent variable in these models. The discussion and justification for the modelling choice here is the same as for the first set of models.

	Variable	Description
	Dependent variables	
1	Firm-authority matches count	ln(#contract awards from public authority to firm in sample)

[1] Repeating the procedure on different samples produces virtually identical results, which can be explained by the very large sample size still remaining after taking the draws - around 1.4 million entries. Also note that the discussion in Adler (2011) is for the case of classification, but an extension to regression follows immediately.

[2] Models which are estimated on the full dataset, containing all 3.3 million transactions, are presented in appendix 2.4. The results are substantively very similar, which is not surprising as these models are capturing the same underlying data generating process. The fit of these models, however, is likely overestimated due to dependence among data points.

2	Firm-authority distance	ln(distance in km between city of firm, city of authority)
	Independent variables	
1	CPV code	317 levels indicating the main three-digit common procurement vocabulary code for product being transacted.
2	Nature of the product	Indicator for services, supplies (physical goods), works.
3	Type of authority	Indicator for: national govt, local govt, utilities, EU institution, international organization, public body, other; national agency, regional agency, not specified.
4	Size of contract award	Recorded price of the contract award (lot) in euros; indicator for the 10 deciles of sample distribution.
5	Framework agreement	Indicator for yes/no.
6	Subcontracting likely	Indicator for yes/no.
7	Procurement agency	Indicator for yes/no.
8	Country	Indicator for EU/EEA+associated country transaction takes place in.
9	Procedure type	Indicator for: open, restricted, accelerated negotiated, accelerated restricted, award without publication of contract notice, competitive dialogue, negotiated without call, negotiated with call.
10	The number of offers	Indicator for five quintiles of distribution and missing.
11	EU funding	Indicator for whether part of the contract funded by EU.
12	Criterion for deciding winner	Indicator for lowest price, most economical offer, missing.
13	Firm award count	ln(# contract awards for firm in sample)
14	Authority award count	ln(# contract awards for authority in sample)

Table 2: Descriptions of dependent and independent variables.

In the following, the predictors used in the models are listed, and table 2 summarizes them. The first group includes structural and economic factors that are not immediately connected to the argument outlined in the theory, and therefore, for the most part, serve as competing explanations.

1. The CPV (common procurement vocabulary) code of the transaction. EU contracting rules ensure that a fine-grained systematic description of the good or service being transacted is available. These codes are hierarchical in terms of detail: the first two digits indicate 46 main areas

such as agricultural products, or construction work, and additional digits provide increasing detail. The level of detail is limited to three digits for the RF models, as in many cases digits beyond this are all zeroes, corresponding to no information being provided. This three-digit CPV code provides 317 unique categories in which the object of the contract award can fall. A strong predictive effect is expected from the variable, as differences between markets in terms of concentration and diversity are likely to be significant.

2. An indicator for whether the good is a service, physical good, or public works project. This complements the CPV variable by helping further classify the nature of the transaction.

3. The type of authority making the acquisition. The data allows eight categories for this variable, with the major distinction being between the central government, local government, and public bodies such as utilities.

4. The size of the contract award, in euros. All else equal, smaller contracts should favor more repetitive pairings, because the same amount of expenditure is now divided among multiple matches. Because of this, models should always include this variable as a control, and moreover, as a robustness check, the main empirical model is also re-run on data which has been weighted by the contract award value.

5. Framework agreements. These are complex procedures, in which an agreement for a possibility of future purchases is made. Future purchases are not counted separately in the data.

6. Subcontracting likely. This indicates whether parts of the contract may be subcontracted.

7. Procurement agency. This indicates whether the buyer is a procurement agency, that is a government organization specialized in procurement, that acquires goods and services on behalf of other government entities.

The following set of predictors includes variables which are useful, to various degrees, for testing the governance and transaction-costs argument presented in the theory.

8. The country the transaction takes place in. This fixed effect captures all country-level predictors of the outcome which are not included in the contract-level model. In order to estimate whether well-governed and developed jurisdictions feature higher average levels of diversity, we can check the distribution of predicted country effects, as well as formally estimating the connection between these country effects and an indicator of the quality of governance.

9. The procedure for publicizing and awarding the contract. There are ten possible procedures available under EU legislation. The sample is dominated by the "open" procedure type, which indicates a regular process in which a call for tenders is publicized and firms are then free to submit bids. A few other possibilities are especially problematic from a governance perspective, especially the awarding without publication and the two accelerated procedures (European Commission 2016, Soreide 2002, Graells 2015, Fazekas et al. 2016).

10. The number of offers. A single bidder or a low number of bidders are seen as indicators of problematic transactions by the EU (European Commission 2016, Fazekas et al. 2016).

12. EU funding. If part of the acquisition is funded through EU contributions, this is indicated in the data. These acquisitions are expected to feature more diverse ties, as they are less likely to be extractive in nature, due to the increased oversight.

13. The criterion for deciding the winner. The distinction here is between a lowest-price winning criterion and various "most economical offer" criteria, indicating the inclusion of quality and fit considerations. Both procedures can be abused, so it is hard to formulate a prior expectation. By ignoring product specifications, it is easy for suboptimal transactions to take place, under the

cover of a low price, but at the same time, quality and fit judgements can be subjective and open to manipulation.

4. Results

Figure 1 shows the variable importance plot for a random forest model in which the dependent variable is the number of interactions. The plot indicates the increase in mean squared error when each of the given variables is removed from the model, in the sense of being transformed into random noise. Higher coefficients here indicate higher explanatory importance and horizontal bars indicate 95% confidence intervals. Due to the very large sample size, all included variables have a statistically significant contribution to the model. However, it is also the case that this criterion is quite weak from a substantive perspective, and effect sizes always have to be taken into account.

Overall, the model explains 39.6% of the variation in the dependent variable, as measured on out-of-sample data.

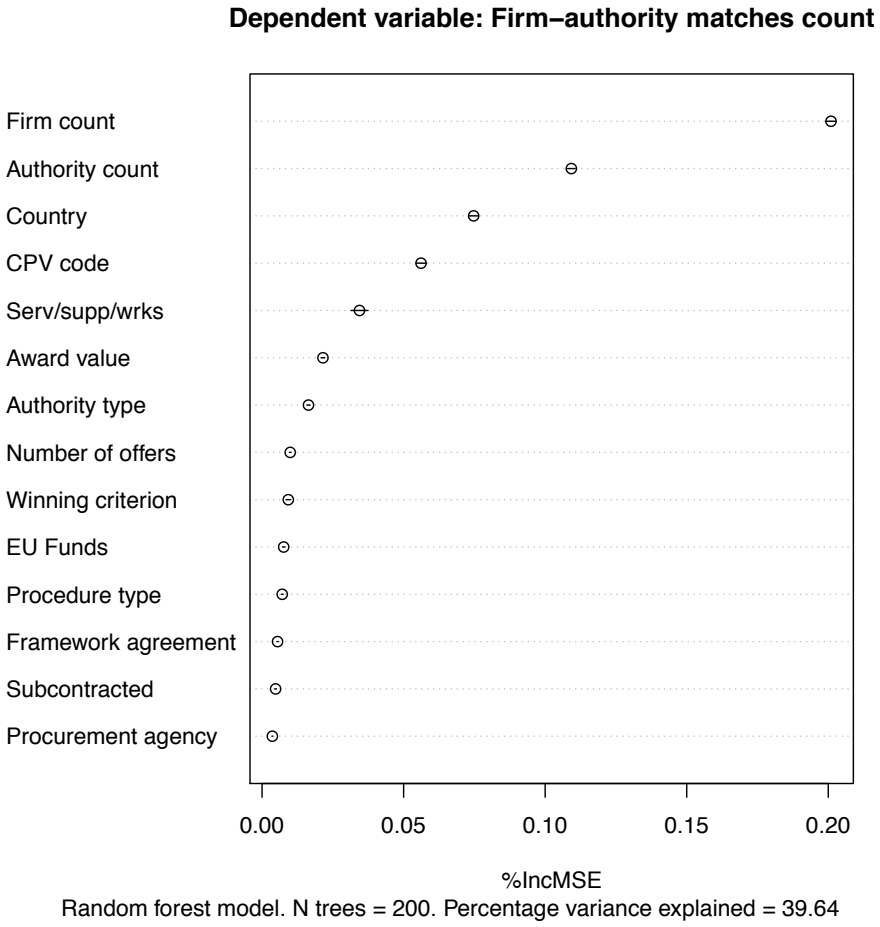


Figure 1: Variable importance plot for transaction-count model

Unsurprisingly, the largest effects on the transaction counts are given by the total contract award counts of the buyer and the seller. The other two major explanatory factors are the country variable and the CPV code of the contract award, while other variables have progressively less explanatory power. In the following we discuss the marginal effects of each of the variables in the model. These are illustrated with an average predictive effects graph obtained by plotting the predicted values generated by the model for each value of the variable, while integrating over the sample distribution of the other variables.

1. CPV code. The overall effect of the variable in the transaction count models, as reflected in figure 1 is very strong. Given the degree of fragmentation of this variable, a full discussion of the patterns emerging among the 317 categories is not practical. However, a sense of its behavior in the model can be had by looking at the most and least diverse CPV codes, as displayed in table 2. (Only CPV codes with more than 100 contract awards in the dataset are included here, to avoid the least substantively relevant ones). The results suggest that, as expected, the least diverse markets tend to be those for high-tech, high fixed-cost, products such as finance, consulting, IT, medicine, and utilities, while the list of high diversity markets generally includes those with a lower technological barrier of entry. As a complement to these results, Hessami (2014) points towards high-tech sectors being associated with corruption in a cross-country setting.

Least diverse	
Banking and investment services	Natural water
R&D and consulting	Forestry services
Computer audit and testing services	Dairy products
Sports services	Prepared and preserved fish
Ships and boats	Agricultural products
Accounting, auditing, fiscal services	Insulated wire and cable
Industry specific software	Training services
Public utilities	Aircraft and spacecraft
Architectural services	Adult and other education services
Installation of medical equipment	Road transport services
	Most diverse

Table 3: Ranking of predicted diversity of ties by CPV-3 code, least to most diverse. Only CPV-3 codes with more than 100 sample entries.

2. Nature of the product. A similar conclusion arises from the services/supplies/works variable. Figure 2 shows that service contract awards predict somewhat higher levels of concentration than for supplies (a difference of .04 log points). This could be due to the more

specialized nature of these markets, as opposed to many physical supplies markets, in which resellers for the same product can generate higher diversity.

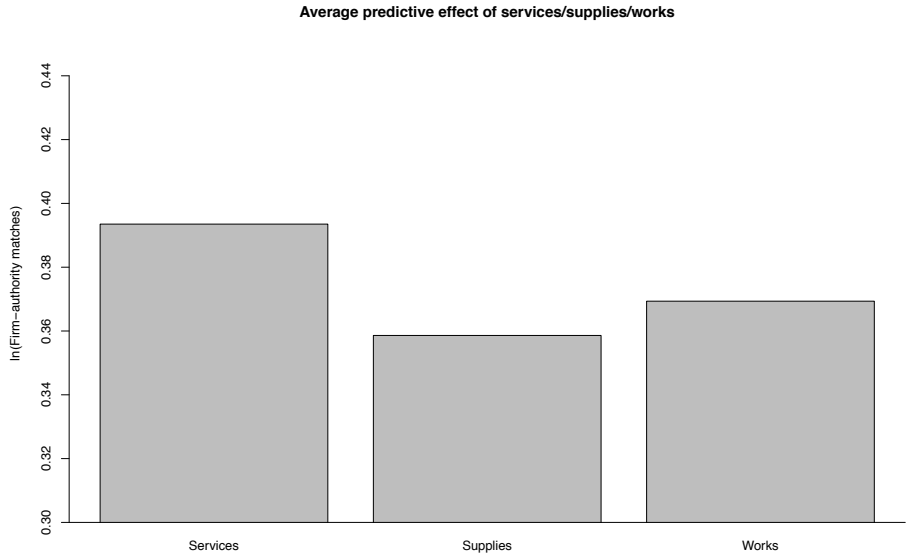


Figure 2: Predictive effect of the type of transaction

3. The type of authority. The data also indicates a reasonably strong effect of the type of authority: local government authorities are somewhat more likely to engage in diverse matches, even after controlling for their likely smaller size, smaller contract awards, and different products. This casts doubt on the idea that local authorities are particularly likely to develop narrow, clientelistic, ties to local firms. Public utilities by contrast show a relatively higher level of concentration. This could receive a number of interpretations: either that they are more prone towards collusive or corrupt behavior, or that the specialized nature of their activities warrants less diverse contracting.

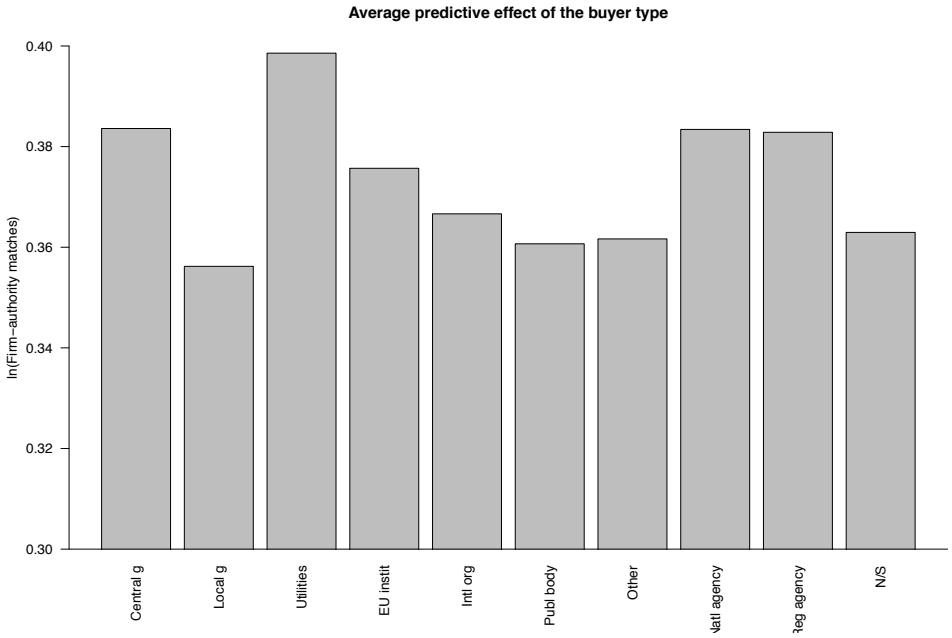


Figure 3: Predictive effect of the authority type

4. The size of the contract award. The effect of the value of the acquisition (figure 4) is as expected. Larger contract awards feature more diverse links, with a difference of .14 log points between the lowest and highest group. One mechanical explanation is that smaller contract awards mean more links being recorded for the same level of expenditure. This will become apparent in the weighted models (appendix 2.2), where the value variable will lose its substantive significance.

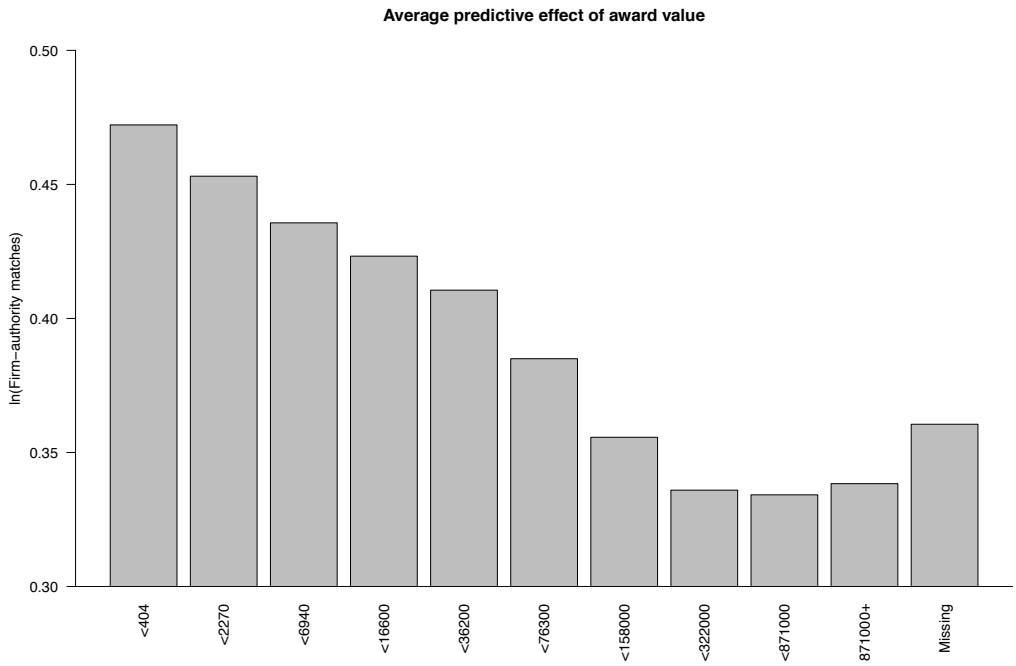


Figure 4: Predictive effect of the award value

The three remaining variables (framework agreements, subcontracting, and the use of a procurement agency), are of a more technical nature and appear to play only a minor role in the predictive model, and will not be analyzed further.

The following variables can be interpreted as evidence towards the validity of the theoretical view connecting socially undesirable outcomes with undiversified ties. The strength of evidence from each of the variables will naturally vary, and the interpretation needs to be commensurately careful.

8. The country indicator.

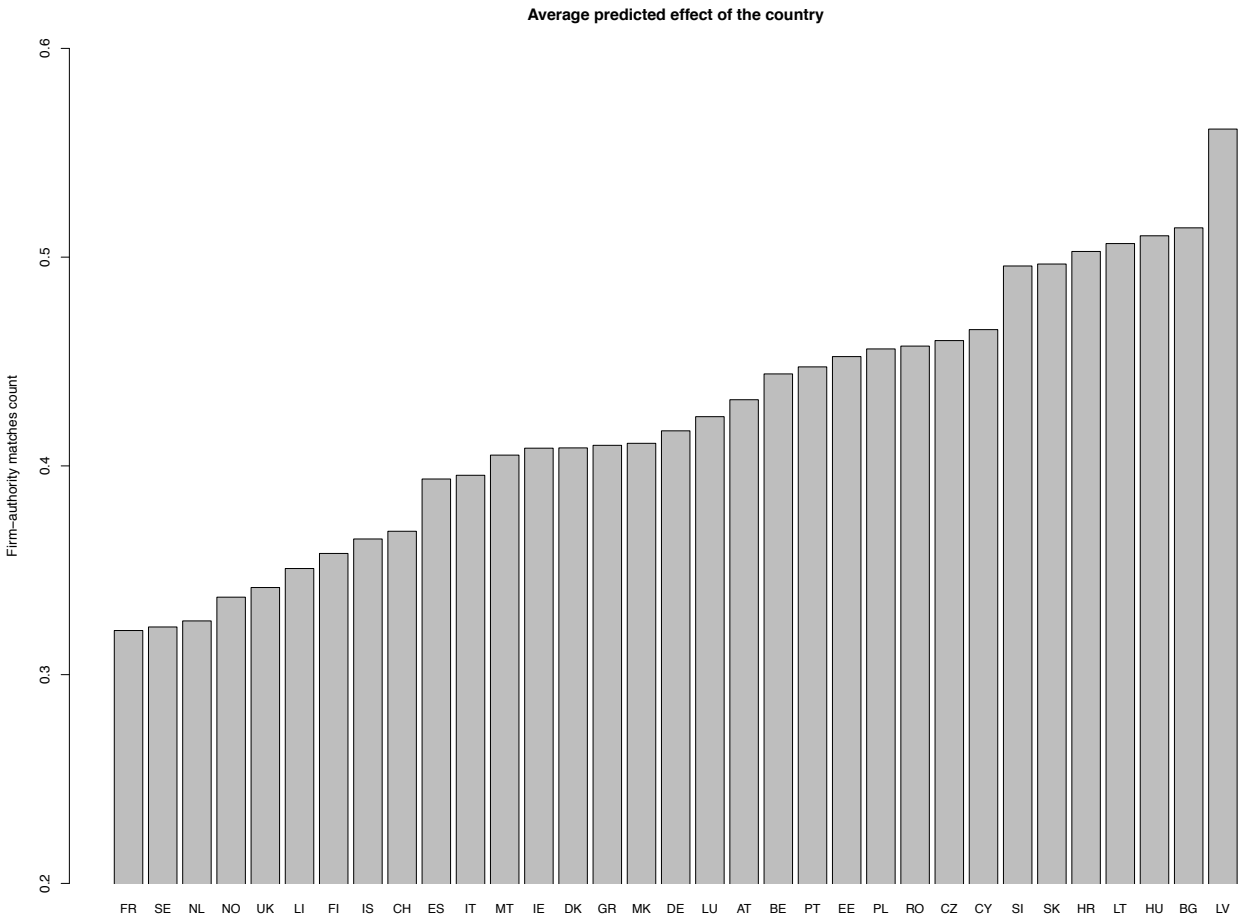


Figure 5: Predictive effect of the country

The effect of the country variable is striking: not only are the differences substantively large (more than .26 log-points between the smallest- and largest-value countries), but the effects map very closely to prior expectations regarding the connection between undiversified ties and environments with poor governance. The intuition can be confirmed with a regression analysis at the country level. The country coefficients from fig. 5 are natural indicators of the prevalent diversity outcomes at the level of each country - they identify the average level of diversity for each country, while keeping the influence of other variables constant. This outcome can be regressed on a widely-used country-level measure of governance quality - the Quality of

Government EQI score from 2013 (Charron et al. 2015), to estimate the effect of the governance environment on average diversity outcomes.[3] For this approach to be valid, we have to assume that any measurement error arising in the country coefficients is random or at least orthogonal to the predictors used in the regression models. If this is the case, then the regression coefficients are unbiased, and any measurement error in the dependent variable translates into larger standard errors on those coefficients (Angrist and Pishke 2008). To protect against heteroskedasticity arising from distribution of the country coefficients, robust standard errors will be used in the regression models.

The predictive effect in the first model is substantively large - moving from the lowest to the highest score predicts an increase of .15 log points in the country average - and strongly significant. Adding a measure of economic development to this model allows to distinguish between the effect of transaction costs arising from governance quality versus simply low income. When adding the logged GDP per capita as a control in model 2, the results are still significant at the .10 level, and the GDP/capita measure is completely non-significant. This suggests that the nature of the process generating the country effects has to do with the governance environment, independent of the level of development. Model 3 adds as a control the natural log of the population, to account for the possibility that larger countries may generate more diverse matches through purely mechanical effects, and this makes the governance variable strongly significant once again.

[3] Note that estimating the diversity - governance relation at the disaggregated level of the RF models would lead to substantial non-independence issues between the data points from the same country, and therefore this approach is avoided.

Country coefficient	M1	M2	M3
Governance	-.044	-.038	-.038
	(.00)	(.09)	(.00)
log(GDP/cap)		-.018	-.021
		(.76)	(.46)
log(Population)			-.019
			(.01)
N	28	28	28
R-squared	.47	.47	.65

Table 4: Linear regressions predicting the country coefficients. P-values in parentheses.

9. The procedure type. This variable has a surprisingly small contribution to the explanatory model, as can also be seen in the variable importance plot in figure 1. The suspicious accelerated and no-publication procedures do not predict less diversified ties (results in appendix 2.1). The variable, by contrast, will have a stronger effect in the distance models to be presented in the next subsection. A possible explanation is that due to the highly suspicious nature of non-open contract procedures, they are avoided by agents intent on misbehaving. Indeed, the crosstab of this variable and the country indicator shows that the poorer-governance new EU member states overwhelmingly use the open procedure for most contracts. If this happens, in equilibrium the variable will not show meaningful connections with other results of poor governance, such as undiversified ties.

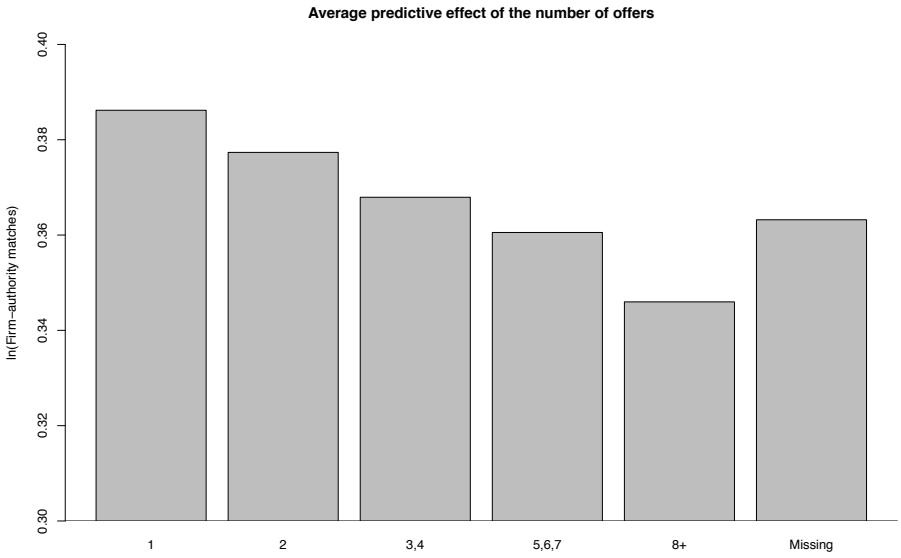


Figure 7: Predictive effect of the number of offers

10. The number of offers. More competitive contract awards predict more diverse ties, with a difference of .03 log points between single-offer and 8-plus offer bids (figure 7). On the one hand, this pattern could simply indicate that lower competition in a market will naturally lead to less diverse pairings, as fewer choices are available for buyers. On the other hand, many structural factors that would determine the competitiveness level, such as the nature of the market, the contract award size, and the total number of transactions for buyer and seller have been controlled for, so what is identified by this variable is competitiveness that is not due to these immediate economic determinants. A low number of bidders is considered an indicator of an inefficient procurement process by both the European Commission (2016), and by the academic literature (Fazekas et al 2016), and is a natural results of a setting in which the existence of a favored supplier is presupposed by market participants. Under this interpretation, these potentially extractive transactions should predict less diversified ties, which is indeed the case in the data.

11. EU funding. The indicator for EU funds has a minor contribution to the explanatory power of the model, but does behave as expected (illustration in appendix 2.1). Projects which are funded by the EU, and are therefore likely subject to more outside scrutiny, do indeed predict slightly more diverse ties.

12. The criterion for deciding the winner. The predictive effect of the criterion for deciding the winner is also illustrated in appendix 2.1, in which lowest-price contracts predict more diversified ties. Given that both options have a theoretical potential to be used for extractive purposes, it is difficult to interpret this finding other than in a descriptive manner.

The following presents results on the geographical models. The presentation is more abbreviated, with only the most important variables being discussed here. The order of the variables and their indices are the same. Figure 8 presents the variable importance plot for these models. The relative importance of the variables is very similar to the first explanatory model, suggesting that the mechanisms at work should be similar, and this will be confirmed by analysis of the individual predictive effects.

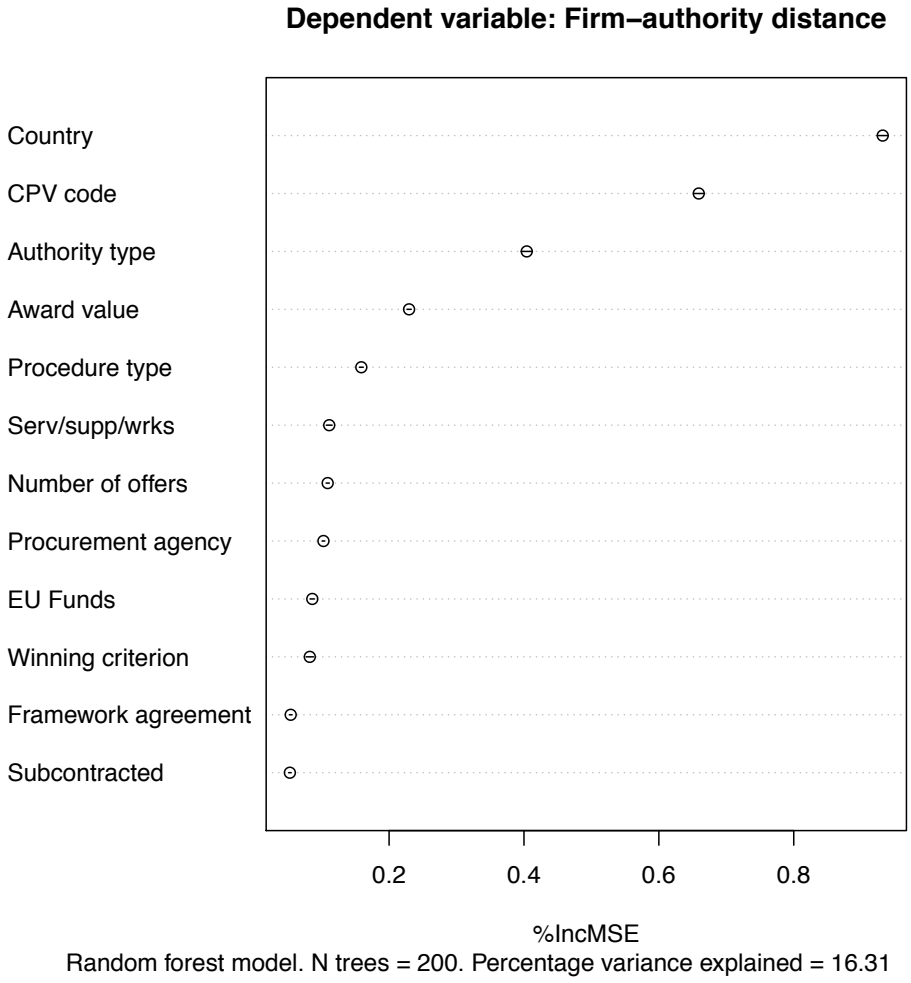


Figure 8: Variable importance for distance model

1. The CPV code. Table 3 displays the largest and smallest predicted values for the CPV indicators. In the case of the distance outcome, the technical characteristics of the market again seem to be very important: while the lowest-distance markets include lower-tech and highly localized services, the long distance markets are generally those for specialized products such as medical equipment.

Closest	...
Primary education services	Basic metals
Real estate services	Luggage
Internet services	Mineral processing and foundry equip
Adult and other education services	Vehicle bodies, trailers
News-agency services	Games, toys, fairground equip
Sporting services	Misc printed matter
Recreational, cultural, services	Medical equipment
Mining equip	Machinery for food processing
Transport services	Pharmaceutical products
Computer equipment and supplies	Misc evaluation or testing equipment
...	Farthest

Table 5: Ranking of predicted buyer-seller distance by CPV-3 code, closest to farthest. Only CPV-3 codes with more than 100 transactions.

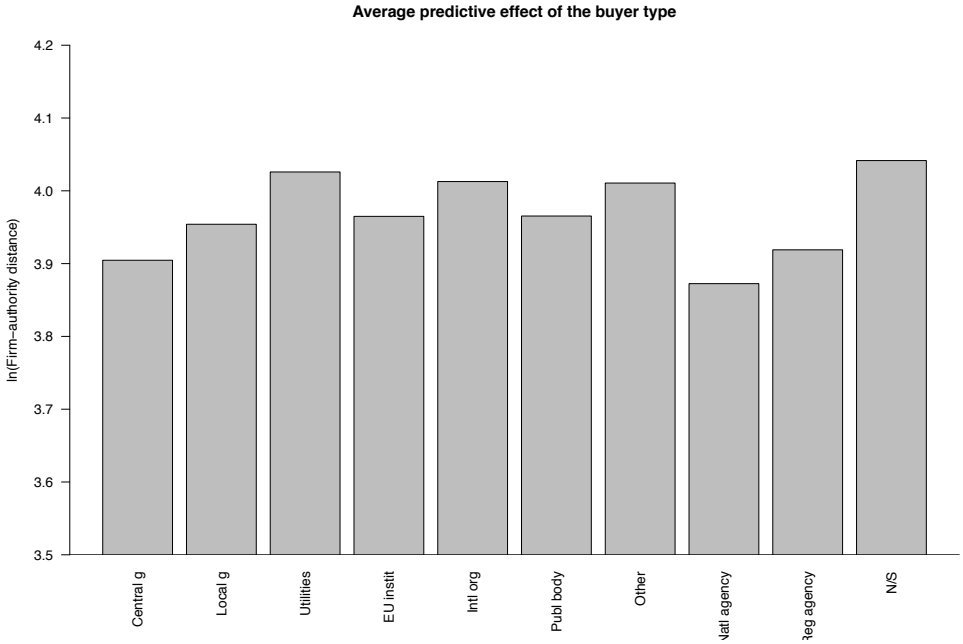


Figure 9: Predictive effect of authority type in distance models

3. The type of authority. Figure 9 shows that central governments and agencies tend to make acquisitions from less distant sellers than either local government or the other types of sellers (a difference of .10 log points between central and local government). This may indicate that much

of the buying by the central government will take place in the capital city, where many suppliers will be located, and shows that a hypothesis according to which local governments develop clientelistic relations with nearby suppliers is not immediately supported by this data.

The next set of predictors can be used to test the logic of the theoretical argument, this time under a geographical interpretation.

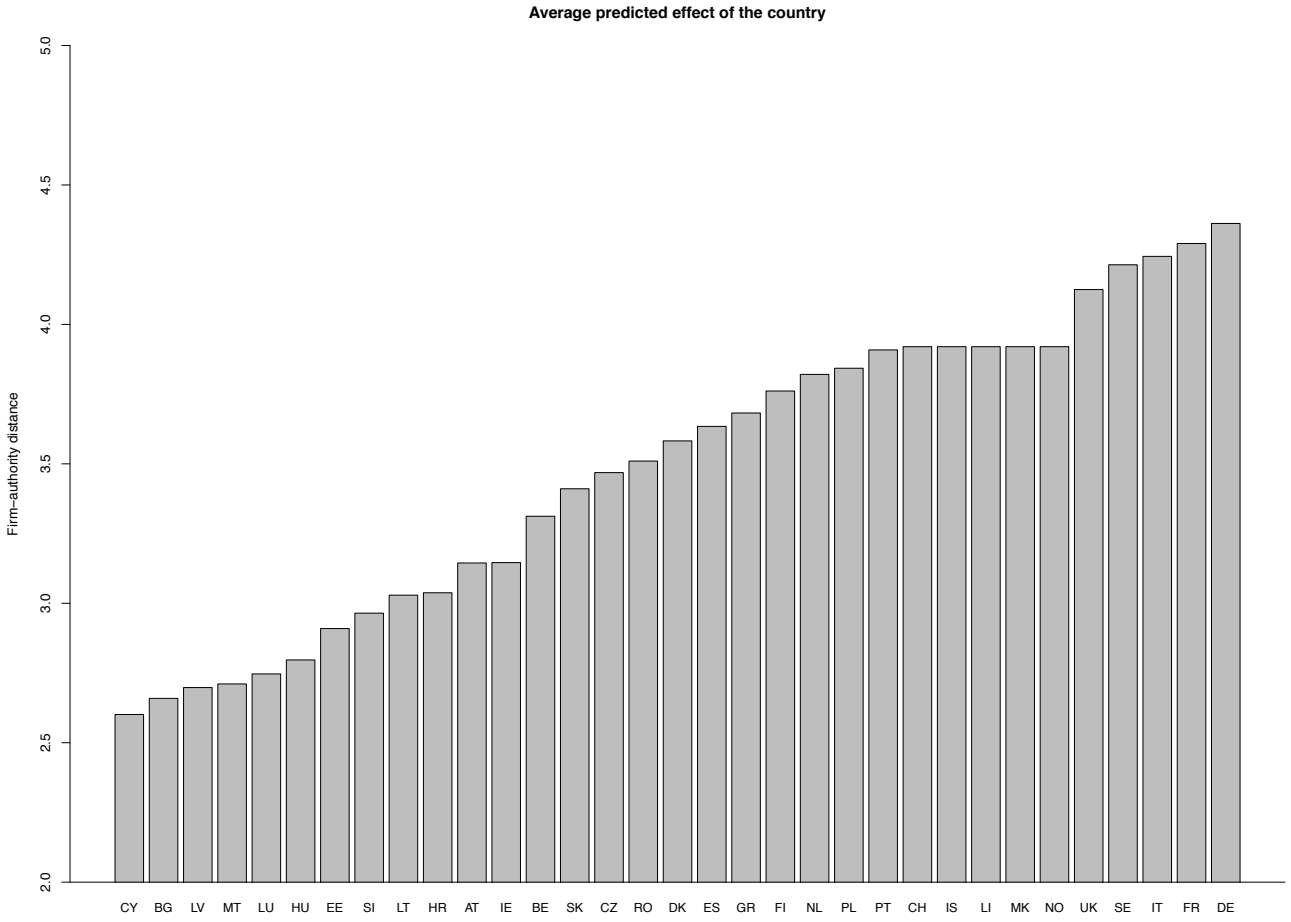


Figure 10: Predictive effect of the country in distance models

8. The country indicator. Interpreting the effect of the country indicators in these models is not as straightforward as in the previous model. A large component of the country effect will be given by the size of the country, which may not be of immediate theoretical interest. However,

even so, the predictive effects in figure 10 are highly suggestive. Table 4 presents results from regression models in which these country coefficients are the dependent variable. To interpret the connection between the governance indicator and predicted geographical diversity, controlling for the area of the country is necessary - while the bivariate model is only marginally significant, once the size is accounted for, the positive connection once again becomes strongly significant. Adding the control for GDP per capita makes the EQI score non-significant, so it is not possible to clearly distinguish between the effects of the two. However, in this case as well, there is evidence of a positive connection between environments with better governance outcomes and geographically more diverse matches.

Country coefficient	M4	M5	M6
Governance	.171	.159	.122
	(.09)	(.02)	(.45)
sqrt(Area)		.002	.002
		(.00)	(.00)
log(GDP/cap)			.119
			(.78)
N	28	28	28
R-squared	.09	.66	.66

Table 6: Linear regressions predicting the country coefficients. P-values in parentheses.

9. The procedure type. Figure 11 shows that the procedure used has a substantively large predictive effect on the distance measure, with a difference of .18 log points between the smallest and the largest predicted value. Unlike in the case of the contract award-count dependent variable, here there is a trend for the less transparent, less competitive, procedures to predict more localized buying. The especially suspicious accelerated, awarded without a call, and restricted procedures are the most localized, while the open procedure is among the most geographically dispersed. The most dispersed procedure, competitive dialogue, tends to be used mostly in the UK. These results suggest that procedures suspected of promoting noncompetitive outcomes are indeed predictive of

a lack of geographic diversity. The difference between these results and the ones for the procedure variable in the acquisition count models, where the procedure variable showed no meaningful patters, is a puzzle. One explanation could be that while the suspicious nature of non-open procedures leads to their avoidance in transactions with favored sellers, this does not affect transactions which are undiversified in the less obvious way of having low geographical diversity.

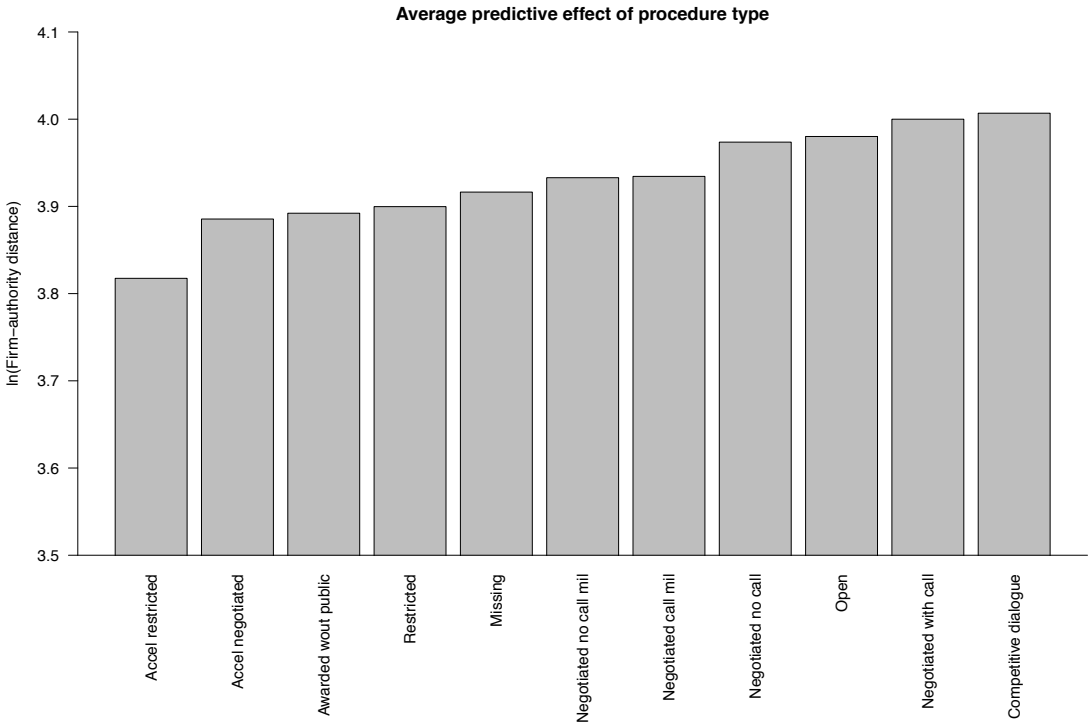


Figure 11: Predictive effect of the procedure type in distance models

10. Competition. Figure 12 shows that a more competitive bidding process predicts longer distances. This again could be because more bidders mean a higher chance for a distant bidder to be selected, but given that many of the structural determinants of distance have been adjusted for, may also be indicative of a process in which inefficient and potentially extractive, transactions predict less geographically diverse ties.

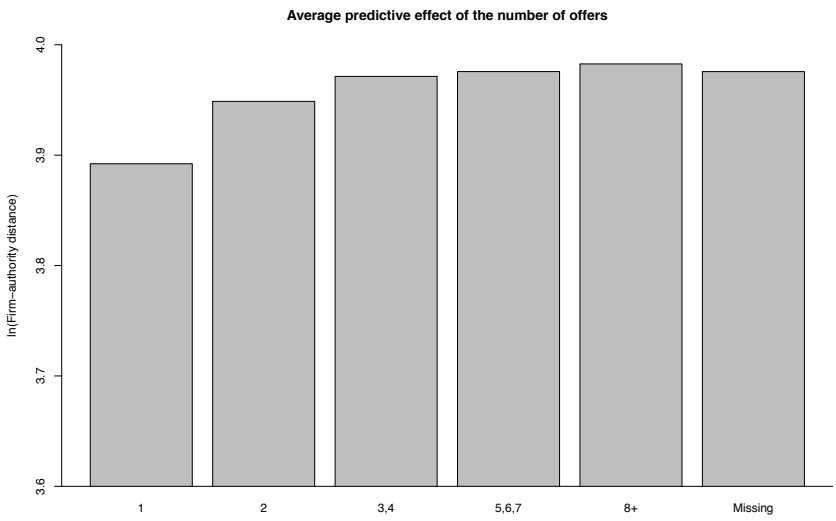


Figure 12: Predictive effect of number of offers in distance models

The other variables again have limited effects so they will not be discussed further.

The appendices present further robustness checks on these results. A first check is given by models in which contract awards are weighted by their value. This is useful in contract award count models as an alternative strategy to controlling for the transaction value in order to ensure like-for-like comparisons. The results from the value weighted models are presented in appendix 2.2, and are very similar to those from the unweighted models. Naturally, in these models, contract award value loses its predictive value for the dependent variable.

A second robustness check comes from considering only contracts with total values above the mandatory-inclusion thresholds, to avoid any potential bias arising from differential inclusion of contracts below the thresholds. (Note that the lower value by itself is not the issue here, as it is controlled for). The results in appendix 5 are very similar in nature to those on the full sample, indicating that any bias arising from this is not substantively important. A discussion for why

identifying the contracts which are truly voluntarily published is difficult is also included in the appendix.

Appendix 2.3 also presents results from the main, unweighted transaction-count models, from samples in which the identities of the firms and authorities are clustered using different cutoff criteria. The variable importance plots of the two supplementary models are almost identical to the results in the body, and the predictive accuracy is slightly lower, as would be expected if more random noise is present. The predictive effect plots for the country indicators are substantively almost identical to the main results as well, as is the case for the other variables (output available in replication materials). From this it can be concluded that the sensitivity-specificity tradeoff of the clustering algorithm does not meaningfully affect the substantive results of the analysis.

Discussion

This article has argued that the structure of matches between public and private actors is indeed connected to governance outcomes. The most important finding is that, even after substantial covariate adjustment, significant differences exist between countries in terms of the predicted diversification of ties between public and private actors, in both contract-count and geographical models, and that these differences are connected to governance quality. This validates the basic theoretical expectation that less diversified ties should be connected to poorer governance. Beyond this, there is some support for the idea that other, transaction-level, indicators of socially undesirable outcomes, which have been previously identified by the literature, such as low competition, non-open contracting procedures, and lack of EU oversight, are also connected to less diversified ties. Taken together, these results suggest that repeated and geographically close

ties between public authorities and firms may emerge when actors are engaging in corrupt or otherwise socially undesirable behavior, and in their turn may favor such undesirable outcomes.

The results also show that some structural and economic features of the contract are connected to undiversified ties. From a practical perspective, the results suggest that contracts for high-tech products, awarded by central governments or utility companies, and of low value, are more prone to the development of undiversified ties. In as much as we believe such undiversified ties then foster inefficient outcomes, this indicates these kinds of contracts should receive increased oversight. Moreover, the findings suggest that geographical proximity behaves in much the same way as repeated interaction for all of these connections.

The conclusions of a line of work on the governance of illicit transactions exemplified by Lambsdorff (2002), and DellaPorta and Vanucci (2004) point in the same direction as this article, but the results here suggest that many questions are still unpursued. How, for example, should we understand the behavior of structural and economic determinants of undiversified ties (such as the nature of the product) with regards to governance outcomes? Are markets which are structurally less diversified more prone to rent generation and outright corruption? Are central governments and public utilities, similarly, more prone to such undesirable outcomes? What is the effect of encouraging procurement from small firms (Kidalov and Snider 2011) on the nature of these ties? Beyond this, important questions regarding the geographical aspect of diversity are arguably still open. We have a very solid theoretical understanding of how repeated interaction reduces transaction costs in settings with weak enforcement, but only an intuitive one of how geography may play the same role, and little empirical evidence to guide us.

A question which is hard to tackle with this data is the extent to which undiversified ties could emerge as socially legitimate adaptations to environments with high transaction costs, in the

absence of extractive, corrupt, behavior. The theoretical section has presented the argument for why this is unlikely, and the results point even more towards this. First, the behavior of the country indicators is hard to make compatible with the connection between undiversified ties and markers such as low competition and suspicious procedures (in the distance models). Far more likely is that the connection arises because settings in which rents are generated through low competition and uncompetitive procedures are also settings in which cooperative behavior between the rent-sharers is facilitated by close ties. Second, in as much as such the transaction costs arise due to reasons other than the desire to hide the nature of the interaction, we would expect them to be connected to income per capita: Less economically developed environments are likely those in which search costs, litigation costs, and other aspects of enforcement are hard to pay for. However, the fact that the country coefficients in our models are closely connected to the governance indicator but not at all to the income per capita measure point away from this mechanism. So, while the possibility of second best efficiency of close ties must be allowed, it is also the case that it is unlikely given these results.

These results encourage a renewed policy focus on the structure of ties between economic agents. Foundational works such as North (1991) and Greif (1993) place the diversification and depersonalization of market interactions at the very center of accounts of economic development. Works on social capital, and the sociological work on weak ties by Granovetter (1977), similarly point to the centrality of this factor. By contrast, applied policy analysis of, in our case, procurement, hardly focuses on this aspect at all: the European Commission's policy analysis papers such as PwC, London Economics, and Ecorys (2011) look in great detail at factors such as the formal rules governing contract awards, but hardly mention the diversity of buyer-seller connections, which, these results suggest, should also be studied carefully. From a policy

perspective, the results here suggest that an important component of institutional reform and anti-corruption drives should be an effort towards diversifying interactions between public and private actors. In the procurement context, this could be done by setting explicit quantitative targets for diversification, as well as by closer auditing of particularly close connections. More generally, ensuring that the same two agents do not have the opportunity to form particularly close connections may be a powerful tool for discouraging and disrupting corrupt interactions.

References

Adler, W., Potapov, S. and Lausen, B., 2011. Classification of repeated measurements data using tree-based ensemble methods. *Computational Statistics*, *26*(2), p.355.

Aggarwal, R.K., Meschke, F. and Wang, T.Y., 2012. Corporate political donations: investment or agency?. *Business and Politics*, *14*(1), pp.1-38.

Angrist, J.D., J.S. Pischke. 2008. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton: Princeton University Press.

Bajari, P. and Tadelis, S., 2001. Incentives versus transaction costs: A theory of procurement contracts. *Rand journal of Economics*, pp.387-407.

Baldi, S., Bottasso, A., Conti, M. and Piccardo, C., 2016. To bid or not to bid: That is the question. *European Journal of Political Economy*, *43*, pp.89-106.

Banerjee, A.V. and Duflo, E., 2000. Reputation effects and the limits of contracting: A study of the Indian software industry. *The Quarterly Journal of Economics*, *115*(3), pp.989-1017.

Boas, T.C., Hidalgo, F.D. and Richardson, N.P., 2014. The spoils of victory: campaign donations and government contracts in Brazil. *The Journal of Politics*, *76*(2), pp.415-429.

Boubakri, Narjess, Omrane Guedhami, Dev Mishra and Walid Saar. 2012. "Political Connections And The Cost Of Equity Capital." *Journal of Corporate Finance* 18(3):541–559.

Breiman, L., 2001. Random forests. *Machine learning*, *45*(1), pp.5-32.

Brown, M., Falk, A. and Fehr, E., 2004. Relational contracts and the nature of market interactions. *Econometrica*, *72*(3), pp.747-780.

Brown, T.L., Potoski, M. and Van Slyke, D.M., 2009. Contracting for complex products. *Journal of Public Administration Research and Theory*, *20*, pp.141-158.

Charron, N., Dahlström, C., Fazekas, M. and Lapuente, V., 2017. Careers, Connections, and Corruption Risks: Investigating the impact of bureaucratic meritocracy on public procurement processes. *The Journal of Politics*, *79*(1), pp.89-104.

Christen, P., 2012. A survey of indexing techniques for scalable record linkage and deduplication. *IEEE transactions on knowledge and data engineering*, *24*(9), pp.1537-1555.

Claessens, Stijn, Erik Feijen and Luc Laeven. 2008. "Political Connections And Preferential Access To Finance: The Role Of Campaign Contributions." *Journal of Financial Economics* 88(3):554–580.

Cohen, W., Ravikumar, P. and Fienberg, S., 2003, August. A comparison of string metrics for matching names and records. In *Kdd workshop on data cleaning and object consolidation* (Vol. 3, pp. 73-78).

Corts, K.S. and Singh, J., 2004. The effect of repeated interaction on contract choice: Evidence from offshore drilling. *Journal of Law, Economics, and Organization*, *20*(1), pp.230-260.

Corts, K.S., 2011. The interaction of implicit and explicit contracts in construction and procurement contracting. *The Journal of Law, Economics, & Organization*, *28*(3), pp.550-568.

Della Porta, D. and Vannucci, A., 2004. The governance mechanisms of corrupt transactions:. In *The new institutional economics of corruption*. Routledge.

European Commission. 2016. 'Single market scoreboard: public procurement'. http://ec.europa.eu/716internal_market/scoreboard/performance_per_policy_area/public_procurement/index_en.htm

Faccio, M. and Parsley, D.C., 2009. Sudden deaths: Taking stock of geographic ties. *Journal of Financial and Quantitative Analysis*, *44*(3), pp.683-718

Fazekas, M. and I.J. Tóth, 2016. 'From corruption to state capture: a new analytical framework with empirical applications from Hungary', *Political Research Quarterly* 69(2): 320–334.

Fazekas, M., Tóth, I.J. and King, L.P., 2016. An Objective Corruption Risk Index Using Public Procurement Data. European Journal on Criminal Policy and Research, 22(3), pp.369-397.

Fazekas, M. and Kocsis, G., 2017. Uncovering high-level corruption: Cross-national objective corruption risk indicators using public procurement data. *British Journal of Political Science*, pp.1-10.

Fisman, R., 2001. Estimating the value of political connections. *The American economic review*, *91*(4), pp.1095-1102.

Goldman, Eitan, Jörg Rocholl and Jongil So. 2009. "Do Politically Connected Boards Affect Firm Value?" *Review of Financial Studies* 22(6):2331–2360

Granovetter, M.S., 1977. The strength of weak ties. *Social networks* (pp. 347-367).

Greif, A., 1993. Contract enforceability and economic institutions in early trade: The Maghribi traders' coalition. *The American economic review*, pp.525-548.

Hansson, L., 2012. The private whistleblower: Defining a new role in the public procurement system. *Business and Politics*, *14*(2), pp.1-26.

Hart, O., and B. Holmstrom (1987): "The Theory of Contracts," in Advances in Economic Theory, Fifth World Congress, ed. by T F. Bewley. Cambridge: Cambridge University Press

Hastie, T., Tibshirani, R. and Friedman, J., 2009. *The elements of statistical learning.* Springer.

Hessami, Z., 2014. Political corruption, public procurement, and budget composition: Theory and evidence from OECD countries. *European Journal of political economy*, *34*, pp.372-389.

Jancsics, D. and Jávor, I., 2012. Corrupt governmental networks. *International Public Management Journal*, *15*(1), pp.62-99.

Jaro, M.A., 1989. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. *Journal of the American Statistical Association*, *84*(406), pp.414-420.

Karpievitch, Y.V., Hill, E.G., Leclerc, A.P., Dabney, A.R. and Almeida, J.S., 2009. An introspective comparison of random forest-based classifiers for the analysis of cluster-correlated data by way of RF++. *PloS one*, *4*(9), p.e7087.

Khwaja, A.I. and Mian, A., 2005. Do lenders favor politically connected firms? Rent provision in an emerging financial market. *The Quarterly Journal of Economics*, *120*(4), pp.1371-1411.

Kidalov, M.V. and Snider, K.F., 2011. US and European public procurement policies for small and medium-sized enterprises (SME): a comparative perspective. *Business and Politics*, *13*(4), pp.1-41.

Kingston, C., 2007. Parochial corruption. *Journal of Economic Behavior & Organization*, *63*(1), pp.73-87.

Klašnja, M., 2015. Corruption and the incumbency disadvantage: theory and evidence. *The Journal of Politics*, *77*(4), pp.928-942.

Klein, B. and Leffler, K.B., 1981. The role of market forces in assuring contractual performance. *Journal of political Economy*, *89*(4), pp.615-641.

Laffont, J.J. and Tirole, J., 1993. *A theory of incentives in procurement and regulation*. MIT Press.

Lambsdorff, J.G. and Teksoz, S.U., 2004. Corrupt relational contracting. *The new institutional economics of corruption*, pp.138-152.

Lambsdorff, J.G., 2002. Making corrupt deals: contracting in the shadow of the law. *Journal of Economic Behavior & Organization*, *48*(3), pp.221-241.

Levin, J. and Tadelis, S., 2010. Contracting for government services: Theory and evidence from US cities. *The Journal of Industrial Economics*, *58*(3), pp.507-541.

MacLeod, W.B., 2007. Reputations, relationships, and contract enforcement. *Journal of economic literature*, *45*(3), pp.595-628.

Rosenbaum, M., Billinger, S. and Stieglitz, N., 2013. Private virtues, public vices: social norms and corruption. *International Journal of Development Issues*, *12*(3), pp.192-212.

Mungiu-Pippidi, A., 2013. Controlling corruption through collective action. *Journal of Democracy*, *24*(1), pp.101-115.

Mungiu-Pippidi, A., 2006. Corruption: Diagnosis and treatment. *Journal of democracy*, *17*(3), pp.86-99.

Murray, C.K., Frijters, P. and Vorster, M., 2015. Give and You Shall Receive: The Emergence of Welfare-Reducing Reciprocity.

North, D.C., 1991. Institutions. *Journal of economic perspectives*, *5*(1), pp.97-112.

Olken, B.A., 2007. Monitoring corruption: evidence from a field experiment in Indonesia. *Journal of political Economy*, *115*(2), pp.200-249.

Pei, M., 2016. *China's crony capitalism*. Harvard University Press.

Portes, A. and Landolt, P., 1996. The downside of social capital. *American Prospect*, (26), pp.18-21.

PwC, London Economics, and Ecorys (2011), 'Public procurement in Europe: cost and effectiveness'.http://ec.europa.eu/internal_market/publicprocurement/docs/modernising_rules/cost-effectiveness_en.pdf

Rey, P. and Salanie, B., 1990. Long-term, short-term and renegotiation: On the value of commitment in contracting. *Econometrica: Journal of the Econometric Society*, pp.597-619.

Rothstein, B., 2011. Anti-corruption: the indirect 'big bang'approach. *Review of International Political Economy*, *18*(2), pp.228-250.

Shapiro, C. and Stiglitz, J.E., 1984. Equilibrium unemployment as a worker discipline device. *The American Economic Review*, *74*(3), pp.433-444.

Spiller, P.T., 2009. 3. An institutional theory of public contracts: regulatory implications. *Regulation, Deregulation, Reregulation: Institutional Perspectives*, p.45.

Tonoyan, V., Strohmeyer, R., Habib, M. and Perlitz, M., 2010. Corruption and entrepreneurship: How formal and informal institutions shape small firm behavior in transition and mature market economies. *Entrepreneurship theory and practice*, *34*(5), pp.803-831.

Winkler, W.E., 1990. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage.

Appendix 1: Summary statistics

Variable	Original dataset name	Mean and (s.d). / categories
Firm-authority matches (log)	log(wincaecount)	.36 (.67)
Firm count (log)	lnwincount	2.46 (2.15)
Authority count (log)	lncaecount	4.85 (1.96)
Firm authority distance (log)	lndist	3.95 (2.13)
Country	ISO_COUNTRY_CODE	33 countries
CPV	CPVthree.ord	317 codes
Authority type	CAE_TYPE	National (.08); Local (.30); Utilities (.06); EU (.00), Int org (.00); Public body (.22); Other (.20); Natl agency (.01), Reg agency (.02), N/S (.06)
Number offers	NUMBER_OFFERS	Five deciles of distribution, plus missing category.
Procedure type	TOP_TYPE	Open(.79), Restricted (.07), others infrequent
Award value	AWARD_VALUE_ EURO_FIN_1	Ten deciles of distribution, plus missing category.
EU funds	B_EU_FUNDS	No (.62); Yes (.08), Missing (.28)
Serv/supp/works	TYPE_OF_CONTRACT	Services (.35); Supplies (.54); Works (.10)
Winning criterion	CRIT_CODE	Lowest price (.29); Most econ (.58); Missing (.11)
Framework agreement	FRA	No (.70); Yes (.29)
Subcontracted	B_SUBCONTRACTED	No (.49); Yes (.07); Missing (.42)
Procurement agency	B_ON_BEHALF	No (.69); Yes (.07); Missing (.22)

Dataset count: 1,467,677 after removing entries with no winner information (generally failed tenders), and sampling one transaction for each firm-authority pair.

Table 1: Summary statistics

Appendix 2: Additional results

2.1 Marginal effect of contract award counts

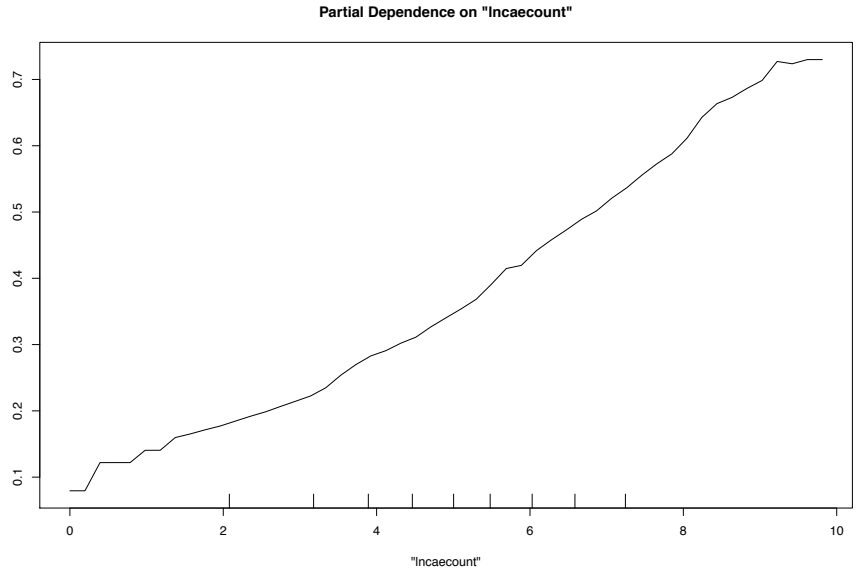


Figure 1: Predictive effect of authority count

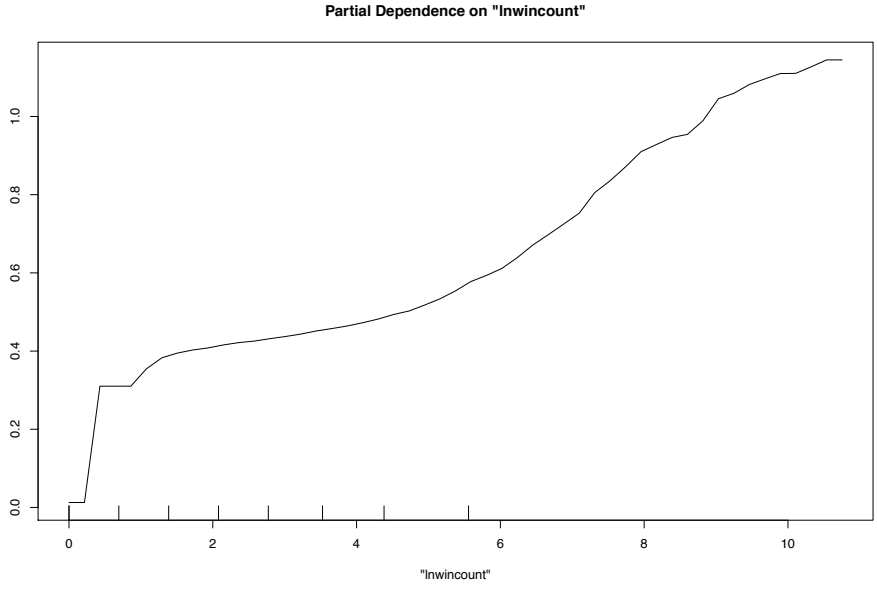


Figure 2: Predictive effect of firm count

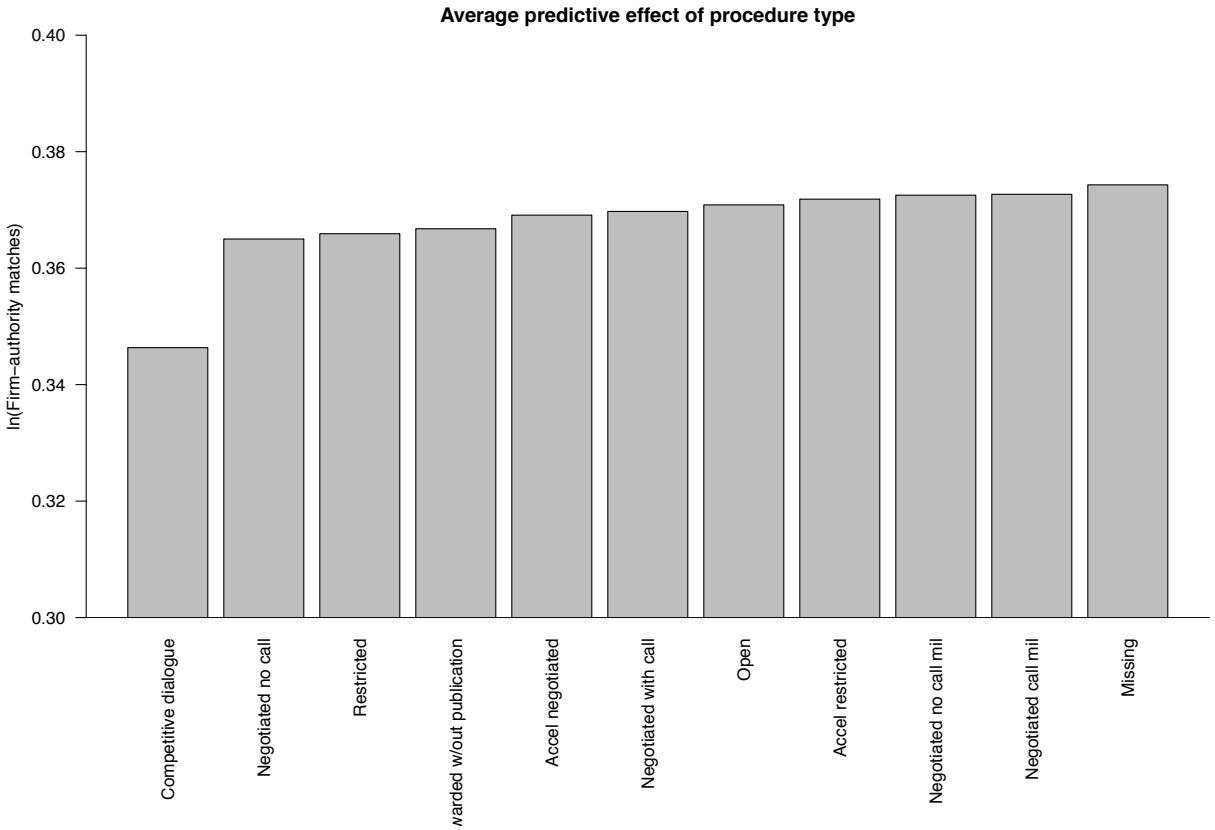


Figure 3: Predictive effect of procedure type

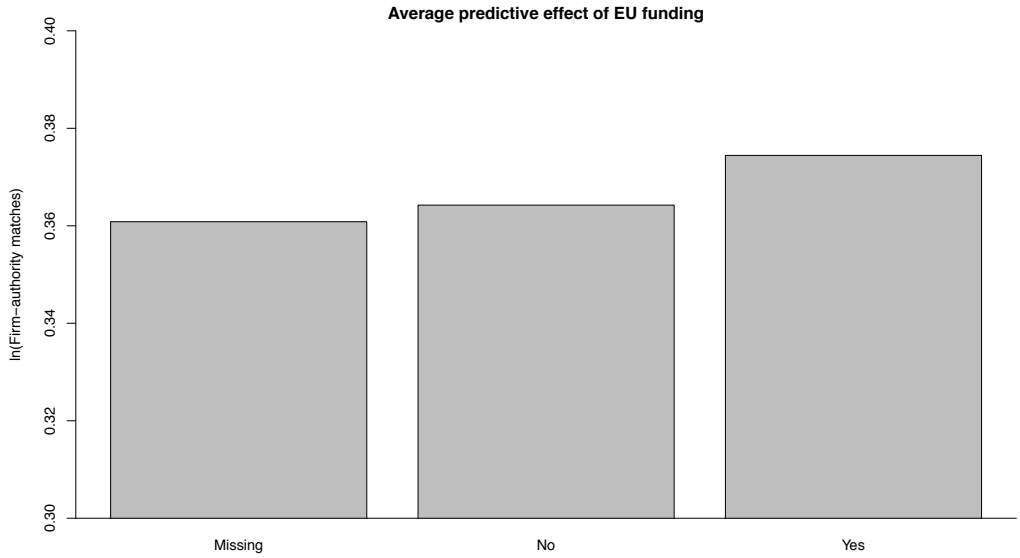


Figure 4: Predictive effect of EU funding

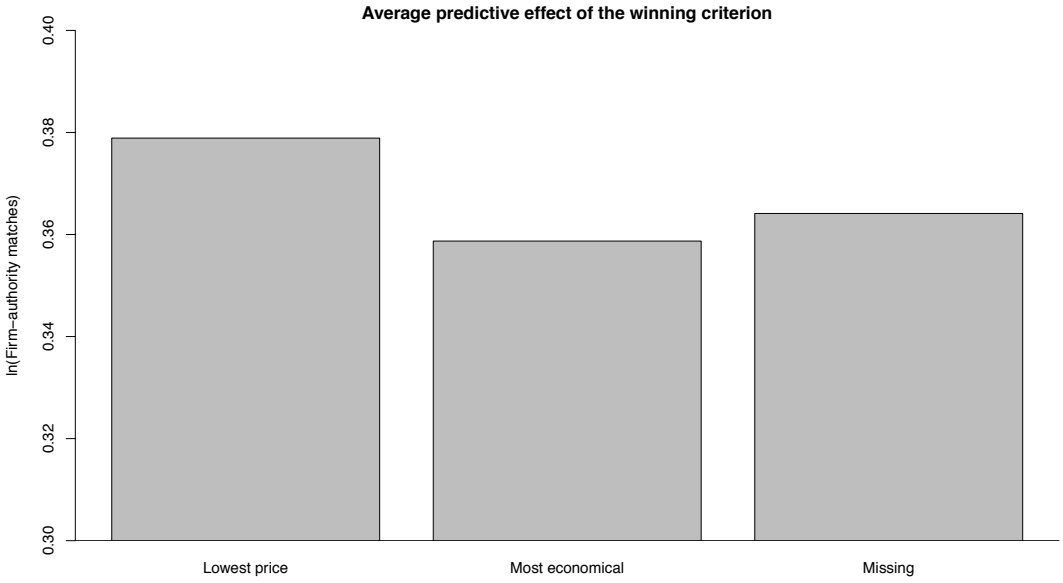


Figure 5: Predictive effect of winning criterion

2.2 Results on weighted models

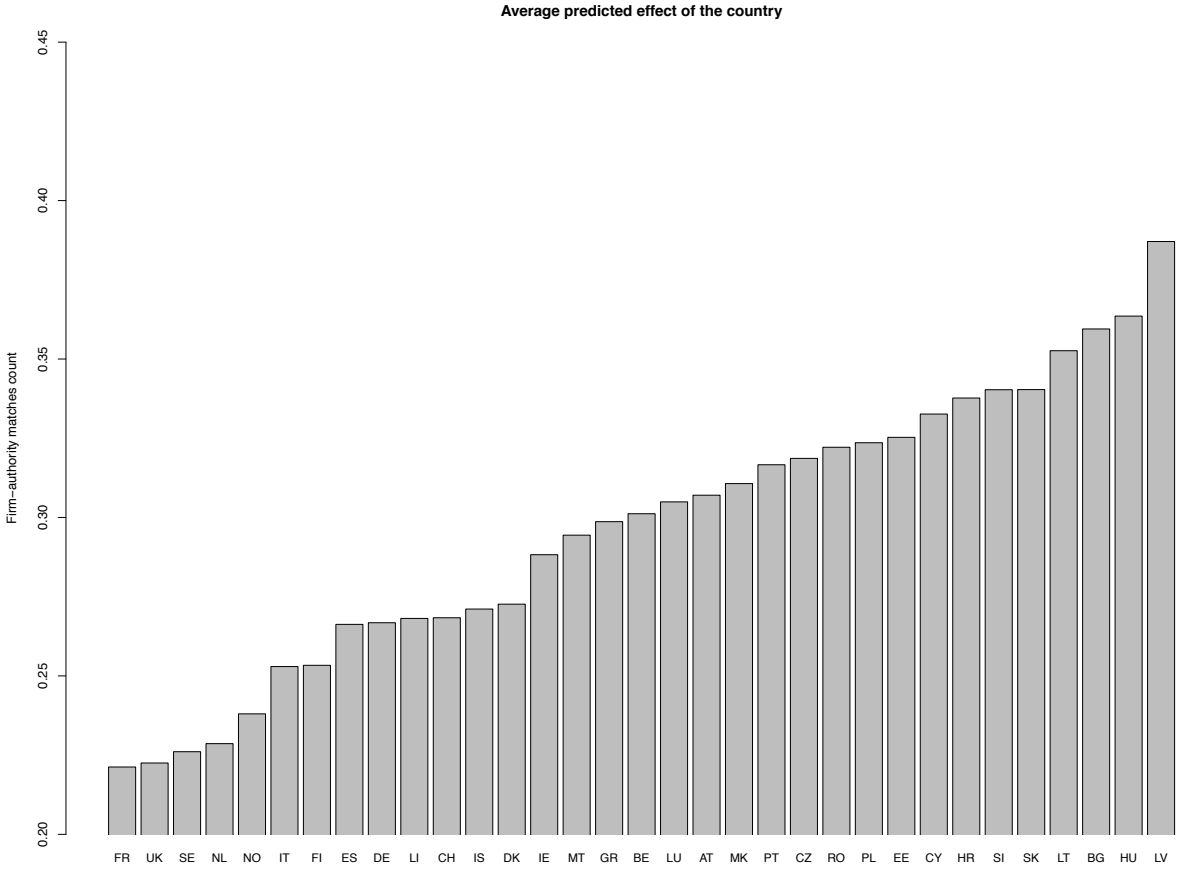


Figure 1: Predictive effect of the country in the weighted model

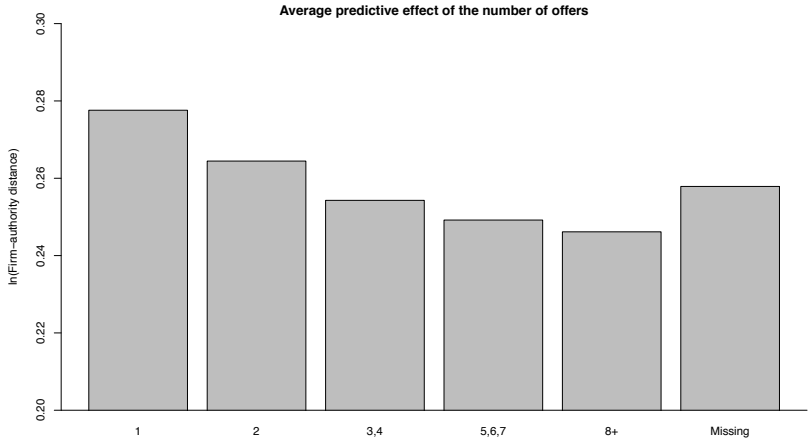


Figure 2: Predictive effect of competition in the weighted model

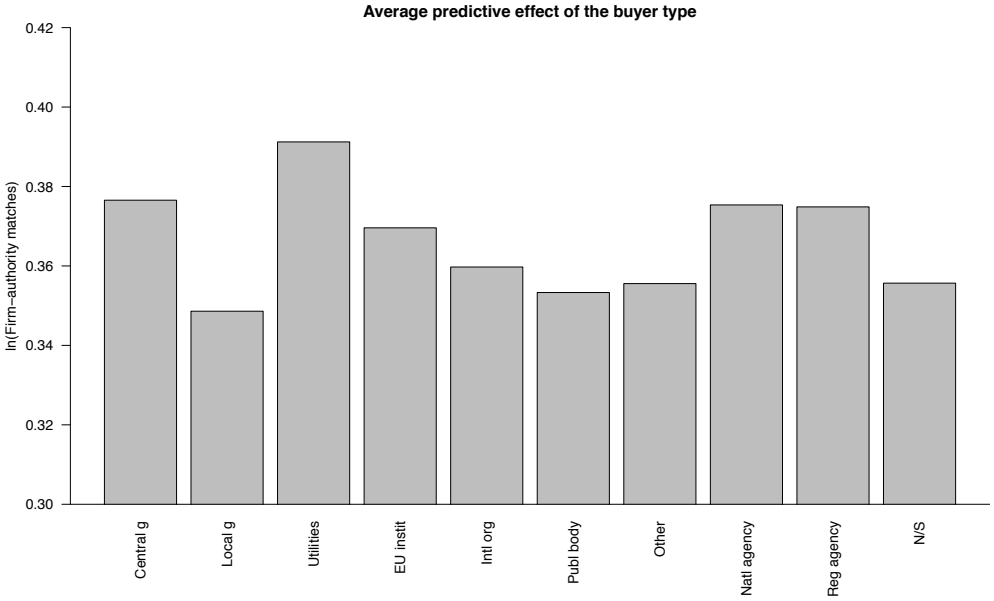


Figure 3: Predictive effect of buyer type in weighted models

2.3 Results with different levels of aggregation in the linkage procedure

2.3.1 Results with firms clustered at the "address-merged" level and authorities at the "cleaned" level.

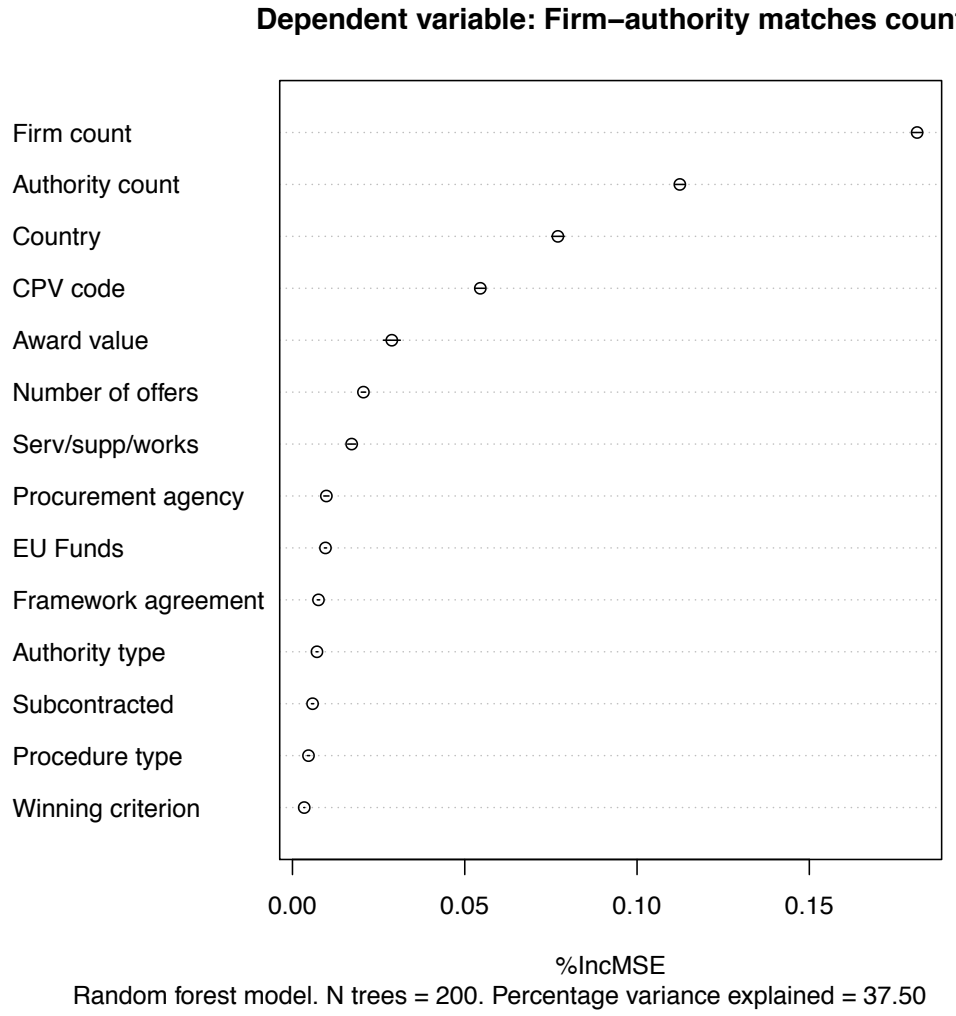


Figure 1: Variable importance plot for unweighted, transaction-count, models with less aggressive record linkage

2.3.2 Results with firms clustered at the .10 level and authorities at the .05 level.

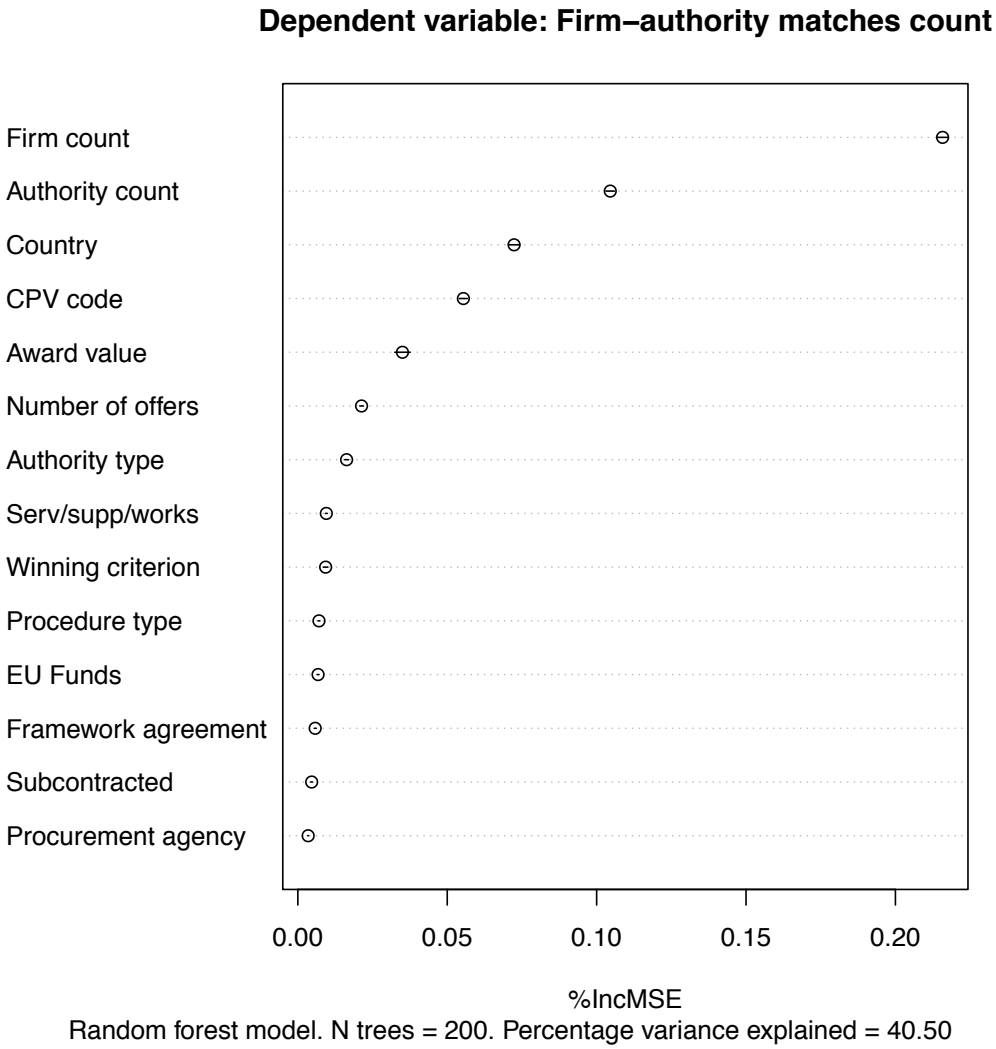


Figure 1: Variable importance plot for unweighted, transaction-count, models with more aggressive record linkage

2.4 Results using data with all transactions per firm-authority pair

2.4.1 Contract count models

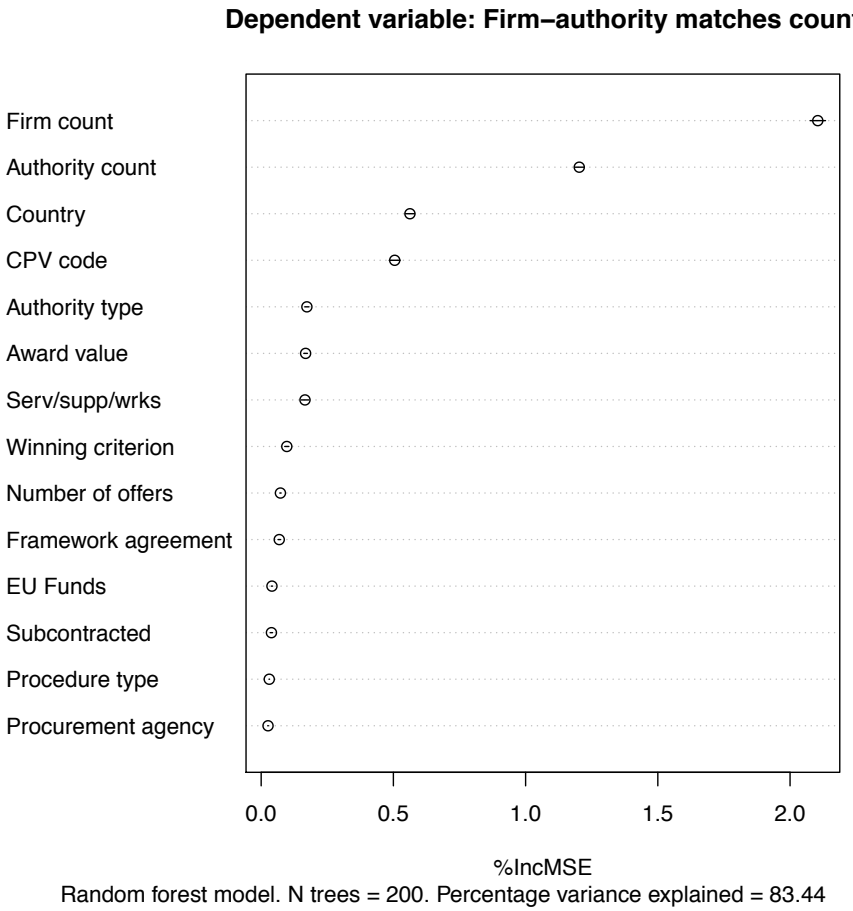


Figure 1: Variable importance plot

Least diverse	
1 Pharmaceutical products	195 Sporting services
2 Aircraft and spacecraft	196 Cybercafé services
3 Drilling services	197 Postcards, greeting cards and other printed
4 Mining equipment	198 Motion picture and video services
5 Medical equipments, pharmaceuticals and personal care	199 Recovered secondary raw materials
6 Medical equipments	200 Apiculture services
7 Textile yarn and thread	201 Equal opportunities consultancy services
8 Fruit, vegetables and related products	202 Animal husbandry services
8 Internet services	203 Machinery for paper or paperboard production
10 Parts of machinery for mining, quarrying, construction	204 Space transport services
	Most diverse

Table 1: Ranking of predicted diversity of ties by CPV-3 code, least to most diverse. Only CPV-3 codes with more than 1000 transactions.

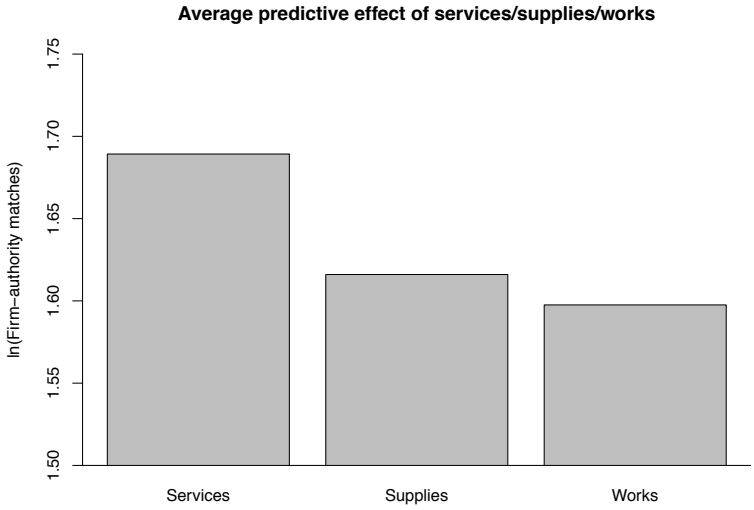


Figure 3: Predictive effect of the type of product

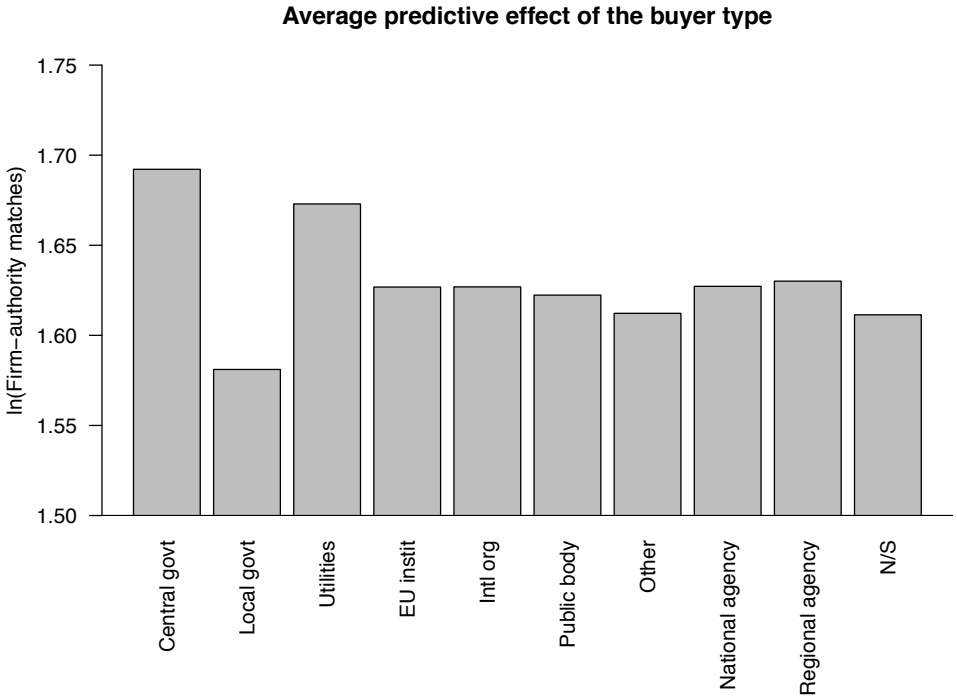


Figure 4: Predictive effect of the type of buyer

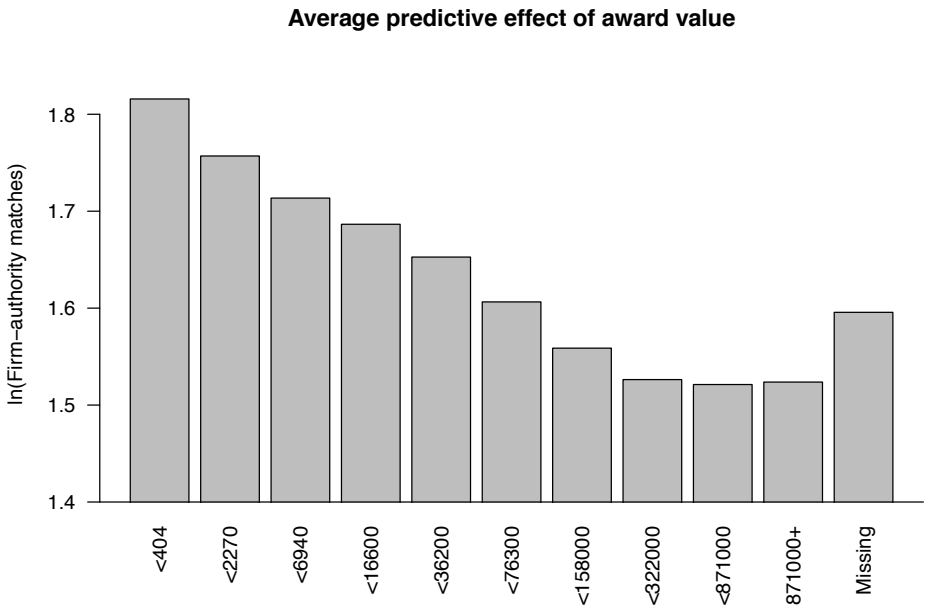


Figure 5: Predictive effect of contract value

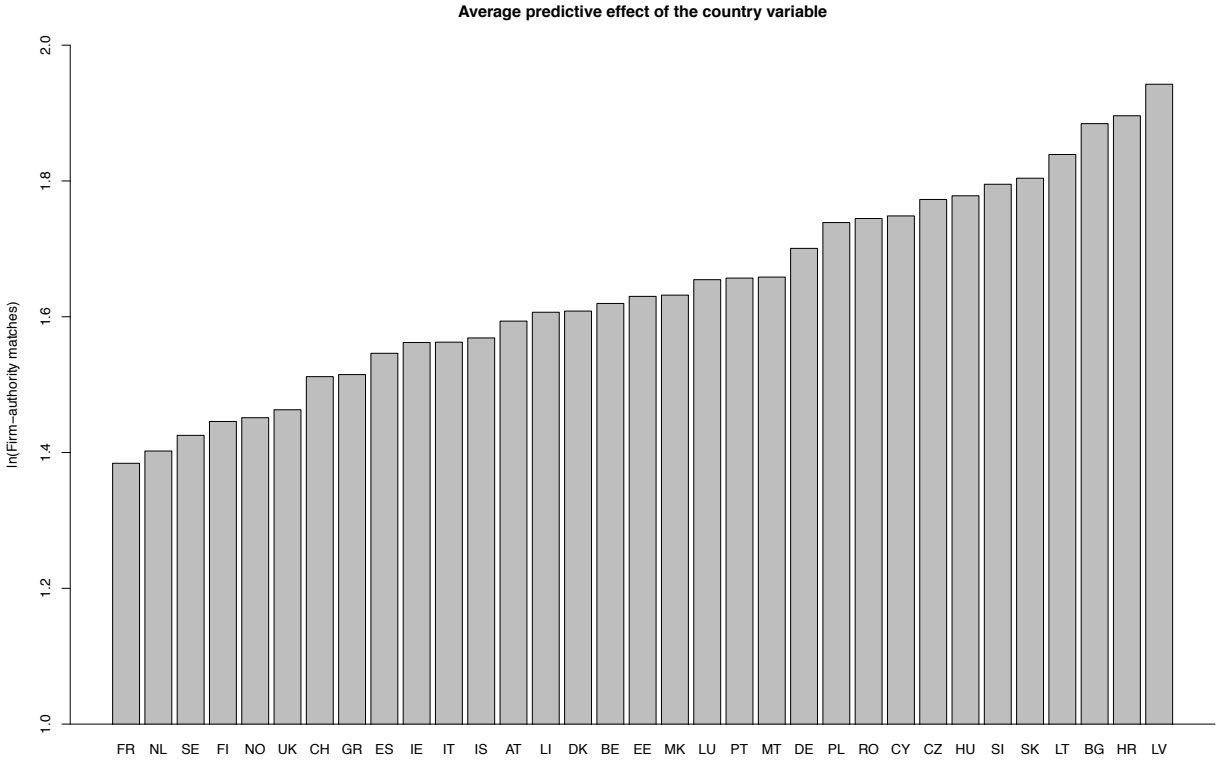


Figure 6: Predictive effect of the country

Country coefficient	M1	M2	M3
Governance	-.110	-.107	-.107
	(.00)	(.02)	(.00)
log(GDP/cap)		-.007	-.107
		(.95)	(.88)
log(Population)			-.044
			(.00)
N	28	28	28
R-squared	.48	.48	.65

Table 2: Linear regressions predicting the country coefficients. P-values in parentheses.

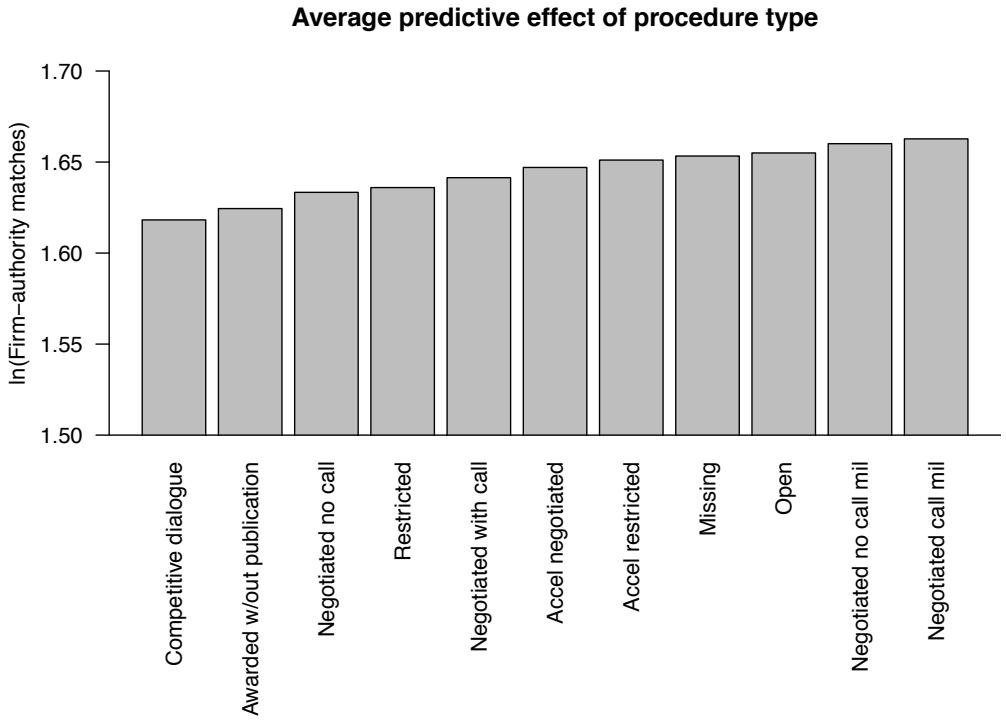


Figure 7: Predictive effect of the procedure type

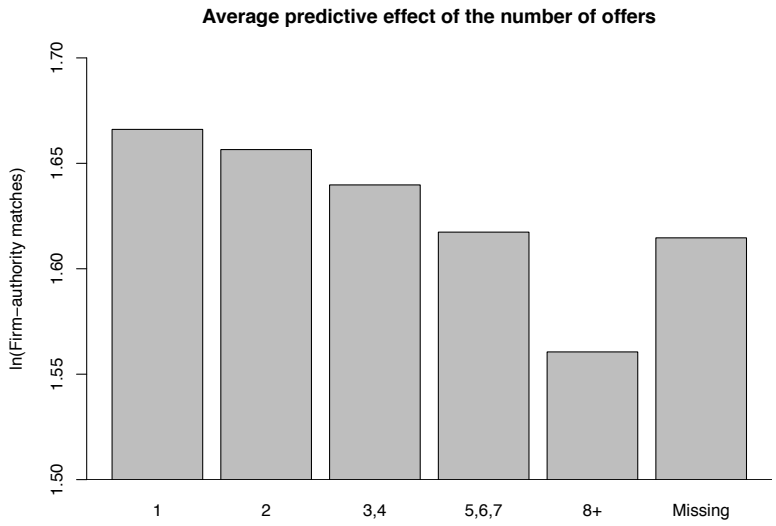


Figure 8: Predictive effect of competition

2.4.2 Distance models

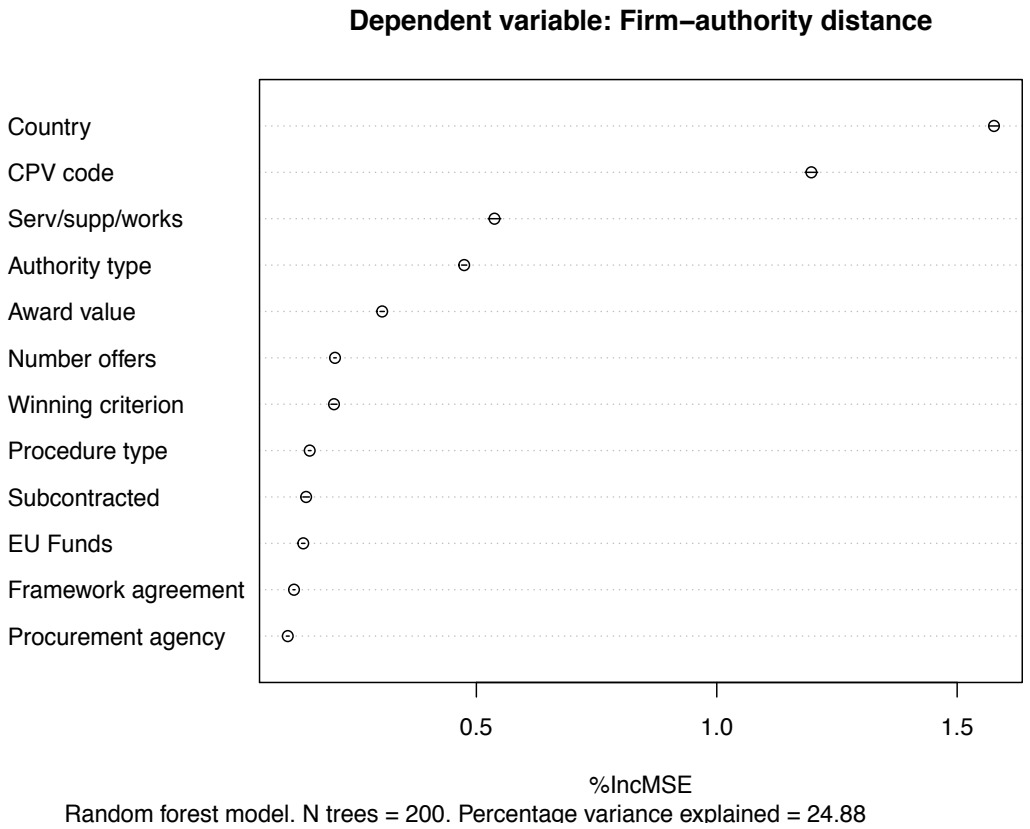


Figure 1: Variable importance plot

Closest	...
1 Agricultural, farming, fishing, forestry and related	195 Medical equipments
2 Miscellaneous equipment (furniture)	196 Lifting and handling equipment and parts
3 Tools, locks, keys, hinges, fasteners, chain and springs	197 Insulated wire and cable
4 Basic inorganic and organic chemicals	198 Special clothing and accessories
5 Research and development services and related	199 Research and development consultancy services
6 Furniture	200 Electricity distribution and related services
7 Petroleum, coal and oil products	201 Electrical machinery, apparatus, equipment
8 Horticultural services	202 Software programming and consultancy services
9 Travel agency, tour operator and tourist assistance	203 Lighting equipment and electric lamps
10 Computer equipment and supplies	204 Refuse and waste related services
...	Farthest

Table 1: Ranking of predicted buyer-seller distance by CPV-3 code, closest to farthest. Only CPV-3 codes with more than 1000 transactions.

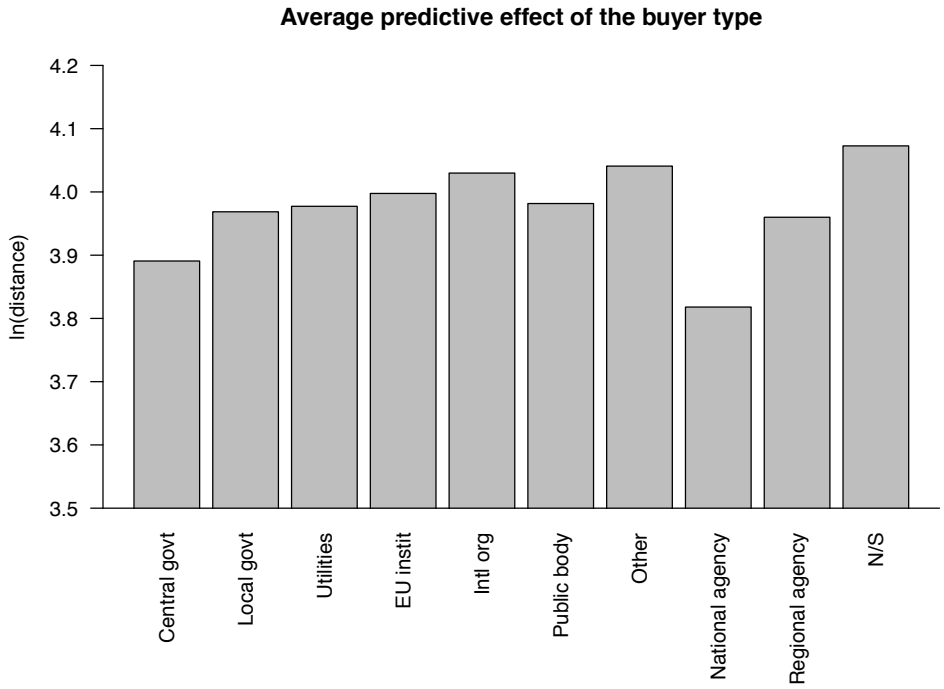


Figure 2: Predictive effect of the type of buyer

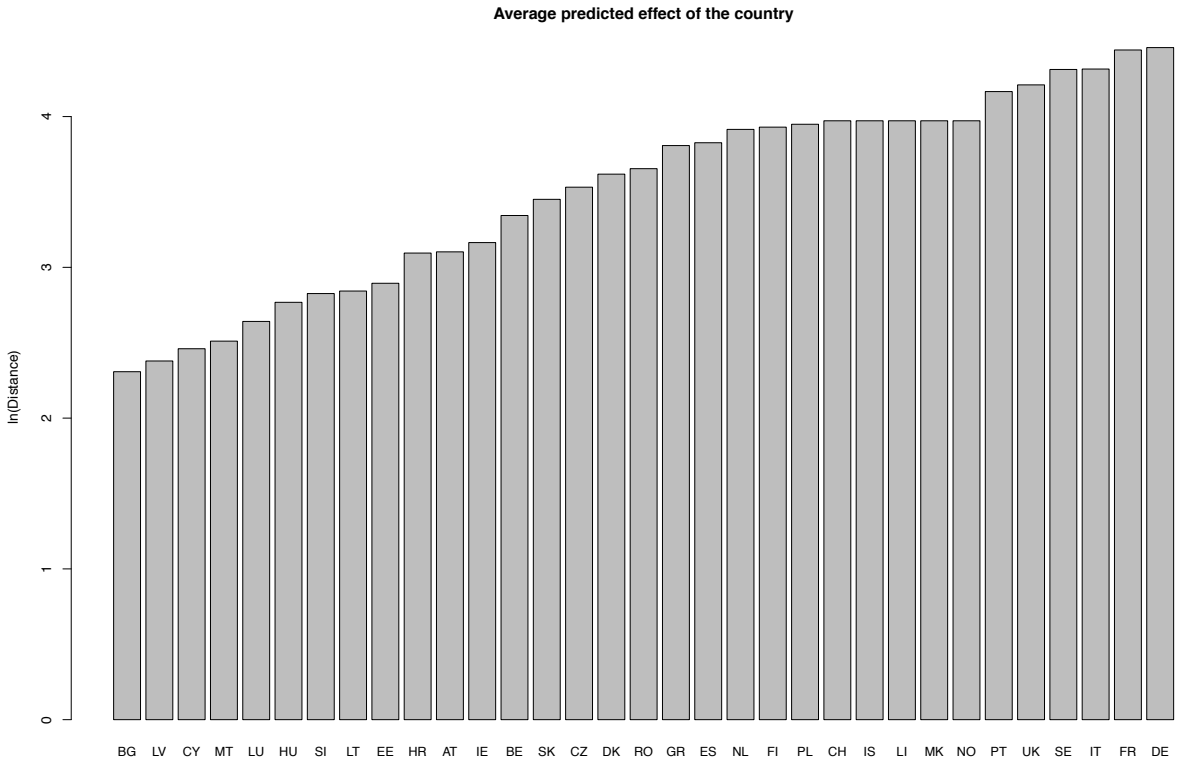


Figure 3: Predictive effect of the country

Country coefficient	M4	M5	M6
Governance	.215	.201	.178
	(.10)	(.00)	(.27)
sqrt(Area)		.002	.002
		(.00)	(.00)
log(GDP/cap)			.071
			(.87)
N	28	28	28
R-squared	.10	.65	.65

Table 2: Linear regressions predicting the country coefficients. P-values in parentheses.

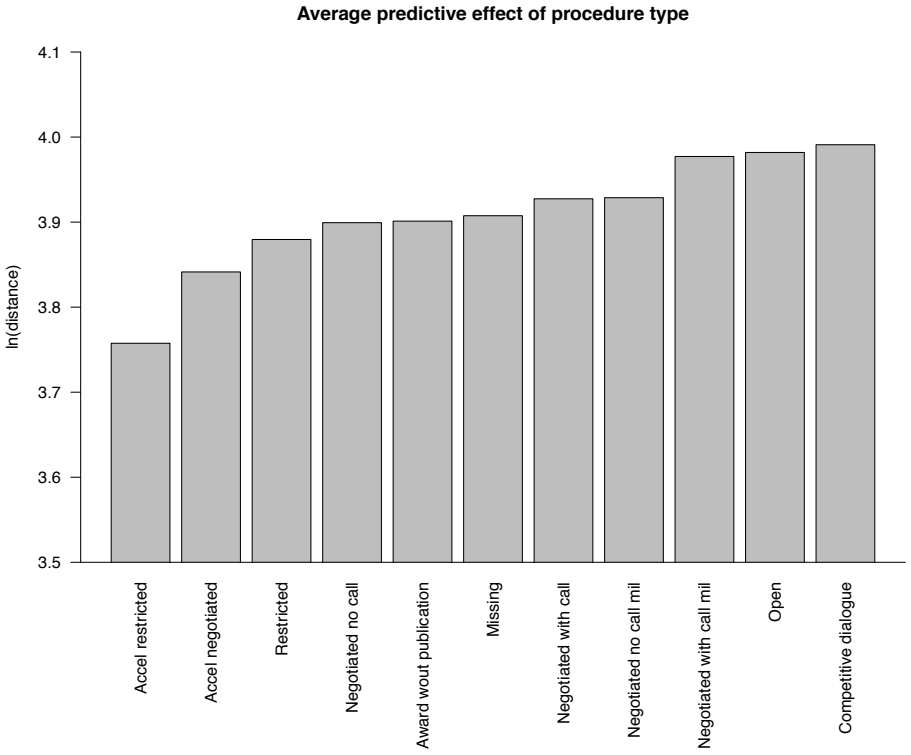


Figure 4: Predictive effect of the procedure type

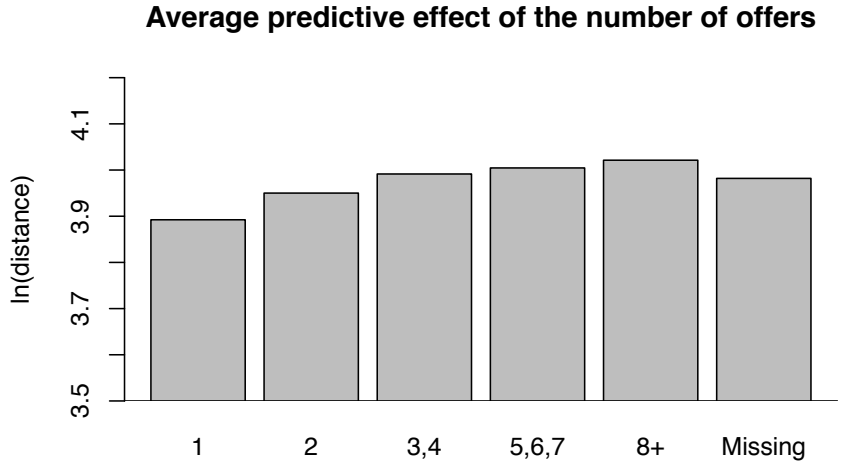


Figure 5: Predictive effect of competition

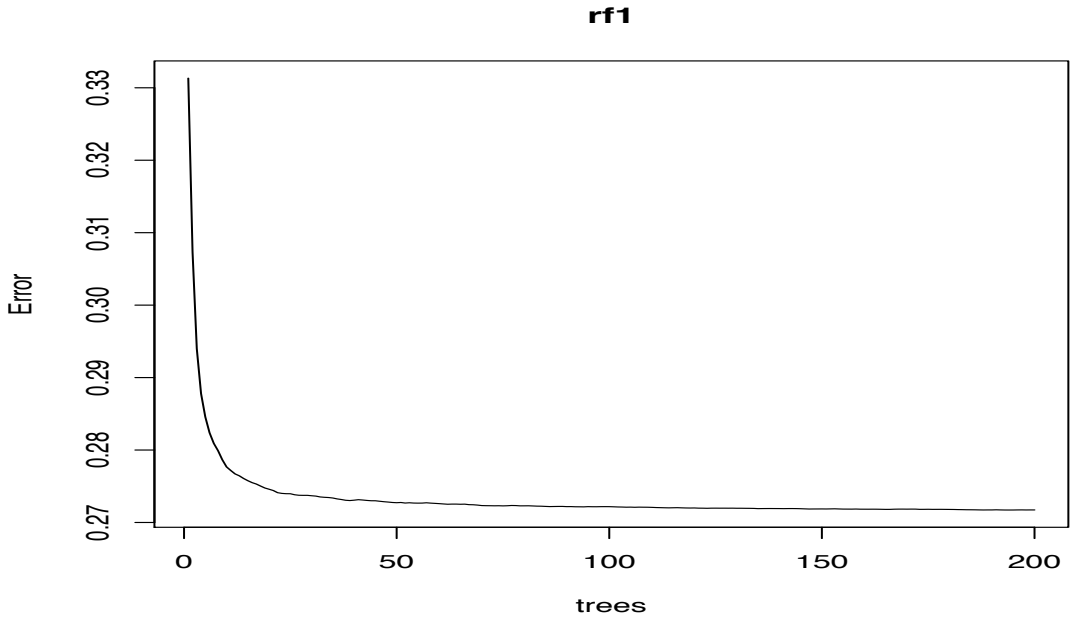
Appendix 3: Statistical methodology

The statistical models are estimated in R, using the RandomForest package (Liaw and Wiener 2002). The command for estimating the main random forest model is:

```
rf1 <- randomForest(y=log(ca$wincaecount), x=ca[,xind], ntree=200, nodesize =
50, sampsize=1/10, proximity = F, importance=T, localImp = F, keep.forest =
T, do.trace=T)
```

Here `ca` is the data frame containing the dataset and `xind` is the vector indicating which predictors to include. In the case of categorical predictors, the default behavior of the random forest algorithm is to search through all possible splits each time the categorical variable is considered. This is impossible to compute for variables with large numbers of categories (there are 2^k possibilities, where k is the number of categories). An insight which greatly simplifies computation for such diversified variables is that in case the outcome is continuous, the categorical variable can be ordered by the average outcome in each category, and then treated as continuous. This will generate the same splits as would the default procedure (Hastie et al. 2008, p 310), so this procedure is used for the CPV variable.

The RF model aggregates 200 individual trees, with convergence in terms of mean squared error achieved after about 50-100. The behavior of the error for the main RF model is presented below:



The node size of 50 allows accurate estimation of the models without becoming overly demanding on computing resources. Setting this value to a lower level does not improve classification accuracy meaningfully, but makes estimation much slower (results available on request). The "mtry" parameter, indicating how many variables to consider at each split is left at its default value, which is the number of variables divided by three. (In practice, the accuracy gain from leaving this at this level, corresponding to estimating a random forest, versus setting it equal to the number of variables, corresponding to estimating a "bagged" model appears to be minimal). The bootstrap sample to be drawn at each step is set at 1/10 of the full sample, so approximately 140,000 data points. This again allows a good balance between accuracy and computational feasibility. Larger samples bring no meaningful increase in accuracy but make estimation more difficult.

The predicted effects plots are computed using commands similar to the following:

```
pp.1 <- partialPlot(rf1, x.var="ISO_COUNTRY_CODE", pred.data=
ca[sample(1:1, 20000))
```

The plots are obtained from random samples of 20000 data points, as there is no need to use the full dataset, which would be prohibitively computationally demanding. It can be checked that drawing repeated samples, or increasing the sample size to larger values does not change the plots in meaningful ways. The plotting command estimates the predicted value of the outcome for each level of the independent variable being plotted. As the other variables need to be kept constant, the quantity is estimated for each combination of sample values, and the results averaged. The procedure is the same as the "average partial effects" obtained in Stata.

The model accuracy (such as the 39% estimated for the first model) is estimated with respect to out-of-sample data. For each draw, the data points which are left out-of-sample are predicted using the full set of trees estimated until then.

An accessible introduction to random forest models is available in James et al (2013), and a more advanced treatment is in Hastie et al (2009).

Appendix 4: Record linkage procedure for firms and authorities

Inspection of the data reveals that the names of the buyers and sellers are not always recorded in a consistent manner. This is to be expected given that the recording is done by potentially thousands of different employees entering the information in the Ted system. While the winning company field requires listing the official name of the entity, this does not preclude a series of problems including inconsistent use of legal designations such as Ltd., Inc., S.A., GmbH, and so on, but also inconsistent recording of the name itself and outright misspellings. In addition, many of the

languages encountered in the sample make use of diacritics, which are difficult to enter consistently and correctly on widely-used English keyboards.

In order to merge the various separate recordings of company and authority names we have broadly followed a procedure which is widely recommended by the statistics and computer science literature on record linkage (Cohen et al 2003), and also implemented in the software packages OpenRefine (Verborgh and De Wilde 2013), and RecordLinkage (Borg and Sariyar 2017). Due to the very large size of the data and various limitations in the packages listed above, we implemented the merging procedure from scratch as will be described. Additionally, we also pursued a few less successful methods, which are briefly discussed.

A recommendation of the literature on record linkage is to acknowledge the probabilistic nature of the process, and rather than searching for the right algorithm to test a series of various procedures, and evaluate each one by drawing a random sample (with a fixed sampling "seed") from the data and checking the accuracy of the merging process by hand. We will do this by randomly selecting 100 contract awards from the full dataset and computing measures of accuracy for various procedures. Additionally, the full R computer code used for the task is make available in the replication materials.

To establish that two names are similar we made use of a measure of string distance. This procedure is common to all record linkage algorithms and solutions, and is based on the idea that while the same entity (firm, person, public institution) may be recorded under slightly different names, they are much more likely to be similar to each other than randomly chosen words.

The distance metric we settled on after some experimentation is the Jaro-Winkler (JW) distance (Jaro 1989, Winker 1990). This is considered especially appropriate for measuring distances between names of entities, as opposed to generic text, and has been shown to have the

best performance among many distance metrics for named entity reconciliation by Cohen et al (2003). The Jaro distance measures the minimum number of character transpositions necessary to turn a string into another, and the JW distance adds Winkler's key insight that often the beginning of the string is more informative than the end. The JW distance uses a parameter ranging between 0 and .25 to give more or less weight to the beginning as opposed to the end of the string. As Winkler (1990) recommends a weight of .10 to be appropriate for most tasks, we also use this weighing. The JW distance between two string ranges between 0 (completely similar) to 1 (completely dissimilar). The Stringdist R package (van der Loo 2016) is used to compute the JW distance, and in order to achieve reasonable execution speeds we employed a remote computing environment on the Microsoft Azure platform.

Step 1: Cleaning the strings.

The first step of all record linkage procedures is a basic "cleaning" of the data. We have therefore performed the following operations on all names of companies:

1. Removing capitalization. This is a standard procedure that is unlikely to affect substantive meaning in any significant way.

2. Removing punctuation. This ensures that, for example, S.A. and SA are the same word.

3. Removing digits. Digits usually appear in the company name field whenever a registration number is included with the company name, or in more unusual situations such as when the address is also mistakenly included.

4. Translating letters with diacritics into their "Latin" counterparts. This is a complex task, which is very well implemented in the Stringi R package (Gagolewski et al 2017). The relevant function is *stri_trans_general*, and the translation is into "Latin-ASCII". This does not affect

words written with Cyrillic characters (from Bulgaria), or with Greek characters (from Greece and Cyprus). This step is very useful given that English Qwerty keyboards are widely used across Europe, making it difficult to input diacritics.

5. Removing the ten most common terms that appear in company names in each country. In all of the countries in the sample, the ten most common terms are designations such as "Inc", "SpA", as well as possibly the name of the country. Such terms are highly unlikely to help differentiate between companies, and are a major source of variation in the recorded names. The full list of terms removed is available in the replication materials.

A similar procedure is employed to clean up the names of authority names. However, we do not remove the most common terms in this case, as they may be substantively meaningful.

In addition, we also cleaned up the winner and authority address fields, by removing capitalization, punctuation, and diacritics. We do not employ more aggressive merging methods for the address fields, as small differences in these strings may correspond to real differences (E.g. 24 Xyz St is very different from 25 Xyz St).

The names of the cities in which the winner and authority are located are cleaned in a similar manner to the addresses, but in addition we also remove any words that are written after a comma or a parenthesis, as sometimes the name of the province is written in this manner.

Step 2: Merging on the address, and on name similarity.

Various experiments with the data have revealed that this is the single most efficient operation for reconciling different recordings of the same entity. While there may be a few ways of writing the same company name, generally the street address is more reliably indicated. (Note that the postal code, city, website, and phone number of the entity are recorded separately.) If two different company names share the same address string, and also have a high degree of similarity,

the it is highly likely that they are the same entity. (Naturally, this should be checked on a sample afterwards.) While this is not a necessary condition for two names to reflect the same entity (consider for example regional offices of the same company), it is arguably a sufficient one.

This criterion is implemented as follows: Inside each country, we consider all names that share the same address string (without the city name). For each such group we record the most frequent name as the label of the group (In case of ties, the first one in alphabetical order is the label.) If a given name is closer than .25 on the JW metric to the label, it is then merged into the label.

This procedure will generate almost no "false positives" (groupings of names that do not belong together), but greatly reduces the name heterogeneity, and ensures that the correct name, in the sense of the most frequently used one, is applied to a large proportion of previously misclassified names (see table 1 in the body of the paper).

Step 3: Clustering on names.

The final step of this and most record linkage procedures is to cluster the names of the entities based on string similarity. This should reflect the idea that names such as "Siemens", "Siemens Corp", "Siemens Healthcare", and so on, belong together on the basis of the fact that they are similar. When dealing with large datasets, however, this can be computationally challenging. To give a sense of the scale of the problem, even after the cleaning the data as described above and merging on the address field, there are still around 180,000 unique names in France, the largest country in the sample. The distance matrix holding the distances between all names will have a size proportional to the square of this number (on the order of 160Gb), and a simple hierarchical clustering procedure would have a length proportional to the third power of this number - and is therefore effectively non-computable even with substantial hardware

resources. To get around this problem we used a "greedy" clustering algorithm, that optimizes locally rather than trying to operate on the overall distance matrix. Such greedy clustering procedures may be less accurate than non-greedy algorithms, but are computable and often provide more than adequate performance. Indeed, table 1 shows that on our sample of contract awards, the algorithm can help achieve above 95% classification accuracy in some configurations, up from around 80% on the unclustered data.

The clustering procedure used is as follows: the names of companies and authorities are sorted in decreasing order of frequency in the data (in case of ties, alphabetical order is applied). For each name, we compute the distance between itself and all names listed before it in the vector of names. (This ensures every name will be compared with every other name in the dataset for each country). If a JW distance under a certain threshold is measured, the two names are merged into the more frequent one. If multiple matches below the threshold are encountered, the closest match is the one that is merged.

The thresholds considered are .05, .10, and .15 for the JW distance. As the accepted distance for a match increases, the rate of false negatives (missed matches) should decrease, but at the same time the rate of false positives (incorrect matches), should increase. In the case of company names, we used the Jaro-Winkler parameter of .10 to give more weight to the first part of the string. In the case of authority names, we have found that it may not be the case generally that the first part of the string is more informative, so the regular Jaro distance (e.g. a parameter of p=0) is used. The R code for this procedure is available in the replication materials.

Evaluation of the algorithm.

To evaluate the success of the various procedures, we draw a sample of contract awards of size 100 from the full data (using the fixed sampling seed 1234). For each of the 100 firms and

100 authorities we record whether it was correctly classified at each step. In order to be correctly classified, an entry has to be both not matched with the wrong label (avoiding a false positive error), and also to not miss any potential matches in the list of names more frequently encountered than itself. Testing the first criterion is easy: for each entry, a judgement can be made on whether the label applied at each step is appropriate (E.g. “huisman muijen adviseur installaties” turning into ““huisman muijen” is appropriate, but “optimare sensorsysteme” into “optimal systems” is an error.) When in doubt, a Google search , together with a translation from Google Translate can be used for this task.

To estimate missed matches (false negatives), we perform a search of the key term or terms for each entry among the full set of names. For example, for “salus international”, we search for all names containing the string salus (even as part of another word). If an entry which we judge to be the same entity is found among those more frequently listed, then the unit fails the false negative criterion. For example, the Irish entry “dhl” was judged as inaccurate, because “dhl express” was also encountered, with a higher frequency.

Table 1 in the body of the paper presents the results of the accuracy test, and reveals three facts regarding the record linkage procedure. The first one is that the non-clustered data is of quite high quality to begin with. Around 79% of companies and 89% of authorities are correctly classified with just a basic cleaning procedure. The second fact is that the merging on address and name step is the most important one for improving classification accuracy: this increases the classification accuracy to 92% for company names and to 97% for authority names. Thirdly, as expected the false positive-false negative tradeoff shifts as the distance is increased in the clustering procedure. Clustering with a .05 distance provides the best balance for company names, but for authority names stopping at the address merging step seems to be optimal. However, as the

various clustering solutions reflect different rates of false positive and false negative error, we will present results with a range of parameters, to show that the basic results do not depend on the precise clustering parameters used.

Unsuccessful record linkage attempts.

In the following we document a few relatively less successful attempts at performing the record linkage, which may be useful to other researchers. The first one was using the API of the OpenCorporates project, which maintains records of registered companies in countries across the world. The API attempts to match given names to companies in the OpenCorporates dataset. However, we found that it was able to match only a small subset (less than half) of our companies, even when those companies not matched are easily located with a Google search. It is hard to say why this procedure fails, but we suspect the fuzzy matching algorithm used by the API is not appropriate. The raw OpenCorporates data is not available to researchers.

The second less successful attempt was to "block" the clustering process on the city of the company or authority, by performing the clustering only inside a city. While this ensures very high accuracy in terms of avoiding false positives, we found it less well-performing than the procedure actually used in terms of avoiding missed matches in all cases.

The third unsuccessful procedure is to use the "textbook" word clustering procedure inside each country, by computing a distance matrix for all names, and then performing hierarchical clustering on that matrix. This only works on the smaller countries in our sample: While sets of up to 30,000 names can be clustered on a desktop computer, as the size increases to around 80,000 (in the case of Germany and the UK), and especially above 100,000 (Poland and France), this becomes computationally infeasible even with high-performance computing resources at our

disposal. As is often the case, however, a simpler greedy clustering algorithm can provide acceptable performance and can be computed much more effectively.

Appendix 5. Results on above-thresholds contracts

The thresholds raise various challenges which make it difficult to identify the contracts which are truly voluntarily published. Contracts are coded as above and below the thresholds depending on whether the contract total value is above the various thresholds which were in place in the years 2009 - 2015, for various types of contracts (central government, local government, and works contracts as a separate category). Doing this tells us that 20.0% of the contract awards for which a contract price was published are potentially under the threshold. This is not a definitive estimate because the publication requirement is based on *estimated* total value, and we only have data on the realized total value. Examination of the distribution of total contract values in the overall sample and in individual countries does not reveal any obvious breaks at around 130,000 or 190,000 euros which are the most relevant thresholds, so it appears that authorities generally do not simply stop publishing contracts that come just under the thresholds. The most significant impediment, however, is that approximately 21% of the contracts do not have the total price data recorded, which makes it difficult to separate those which are under the threshold.

5.1 Results on contract-count models

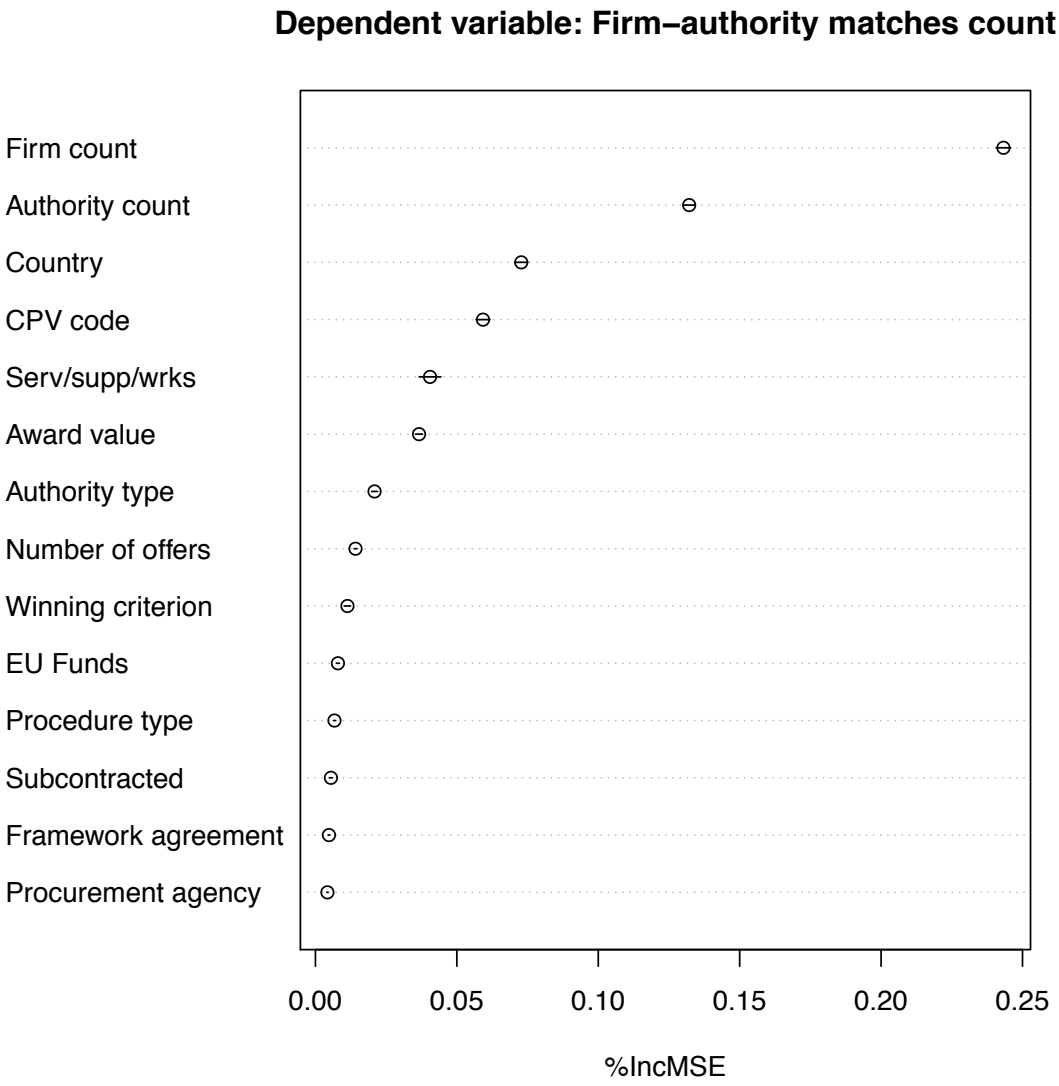


Figure 1: Variable importance plot

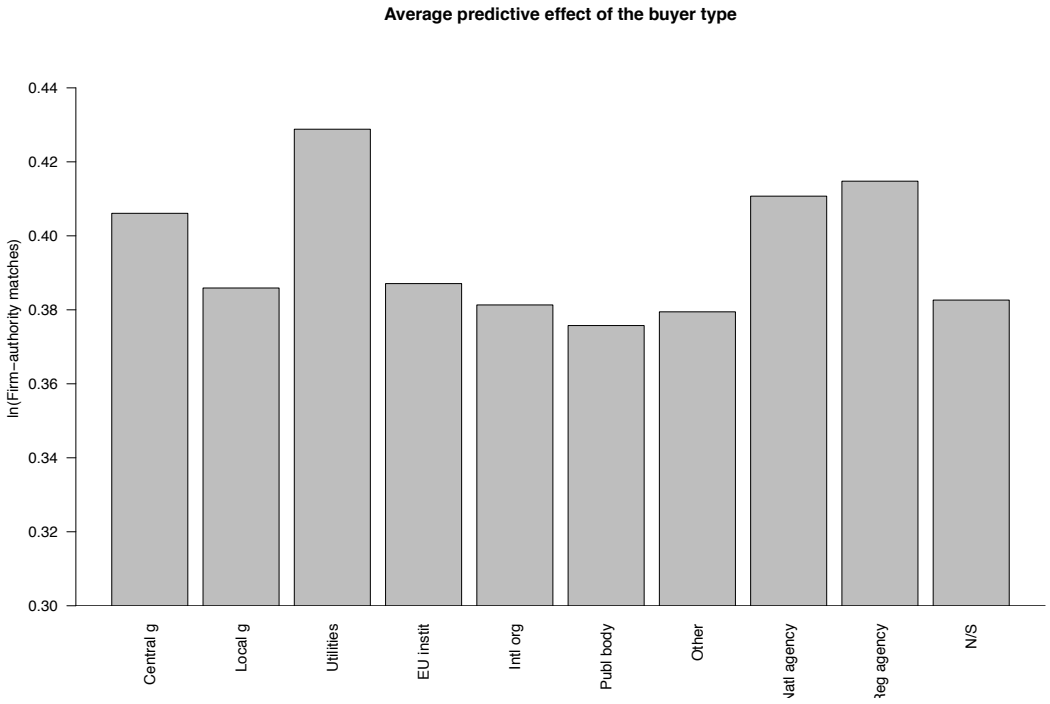


Figure 2: Predictive effect of the type of buyer

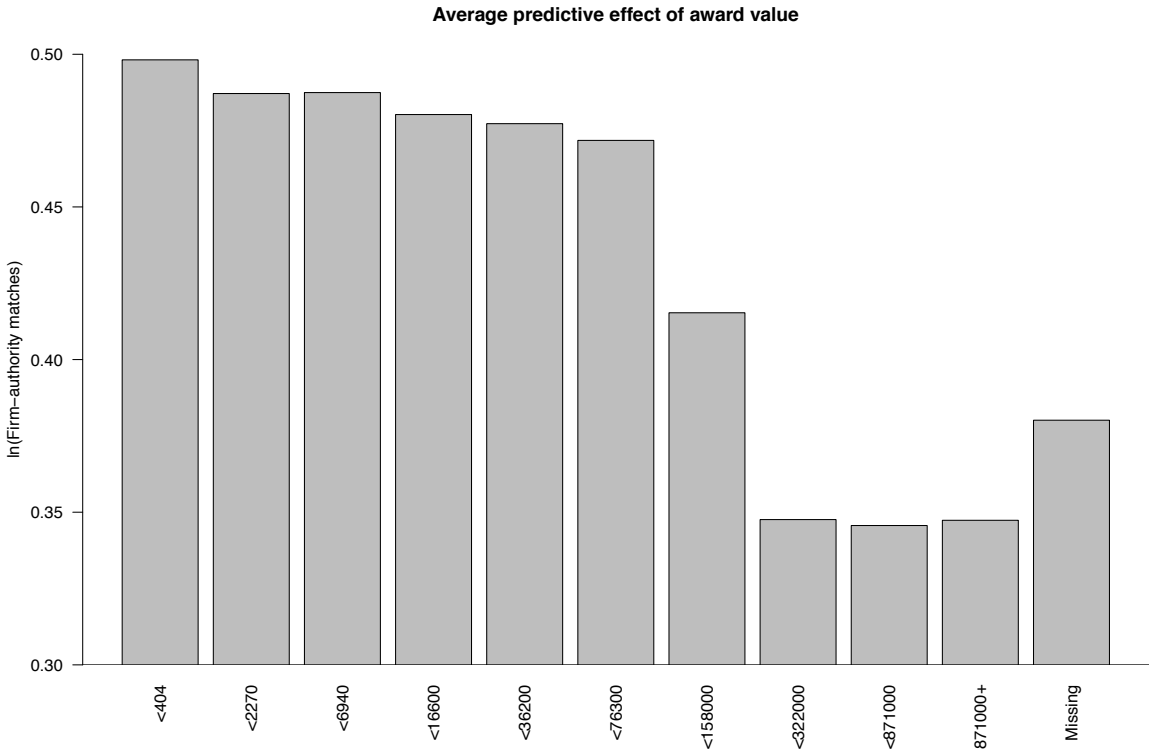


Figure 3: Predictive effect of transaction value

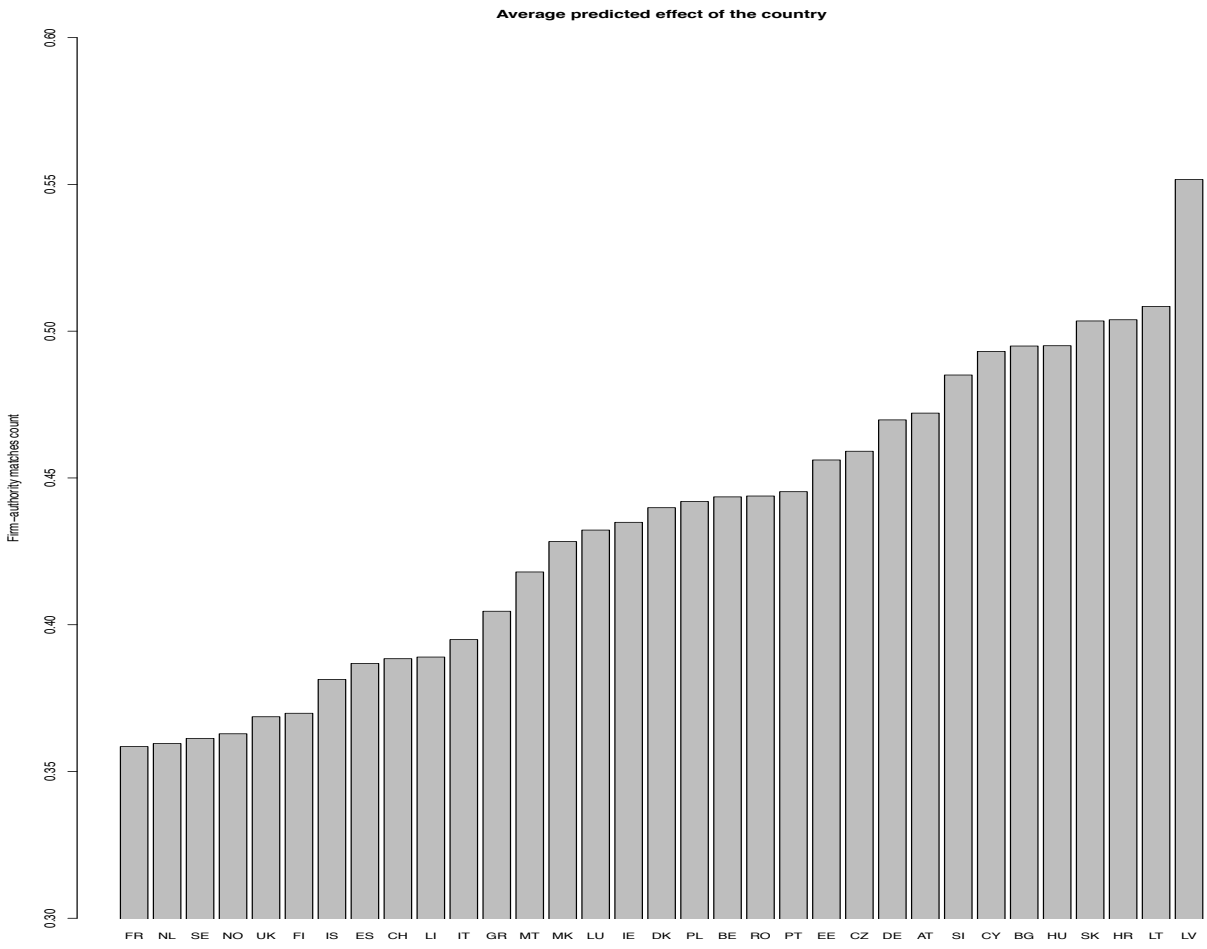


Figure 4: Predictive effect of the country

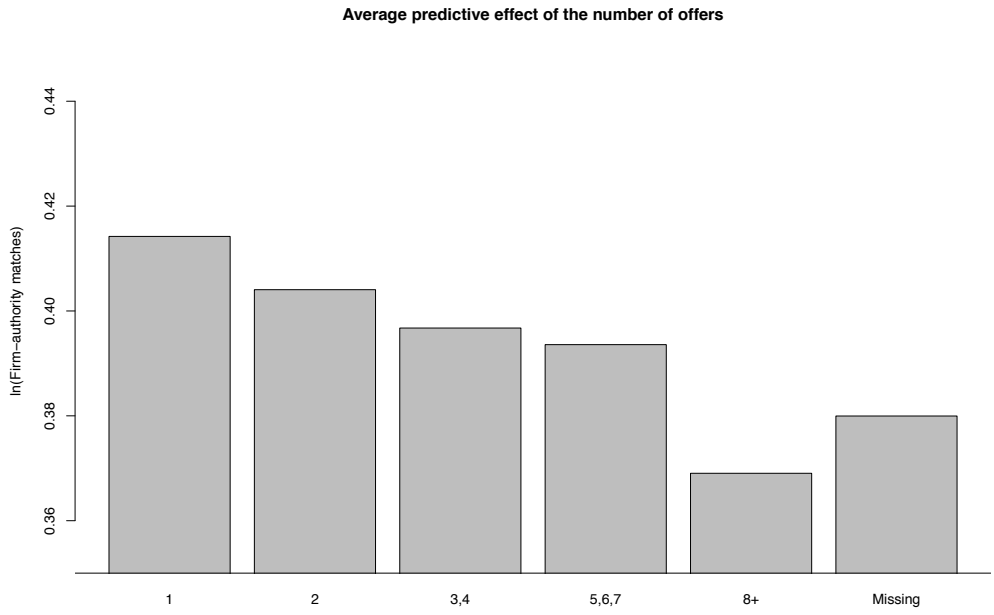


Figure 5: Predictive effect of competition

5.2 Results on distance data

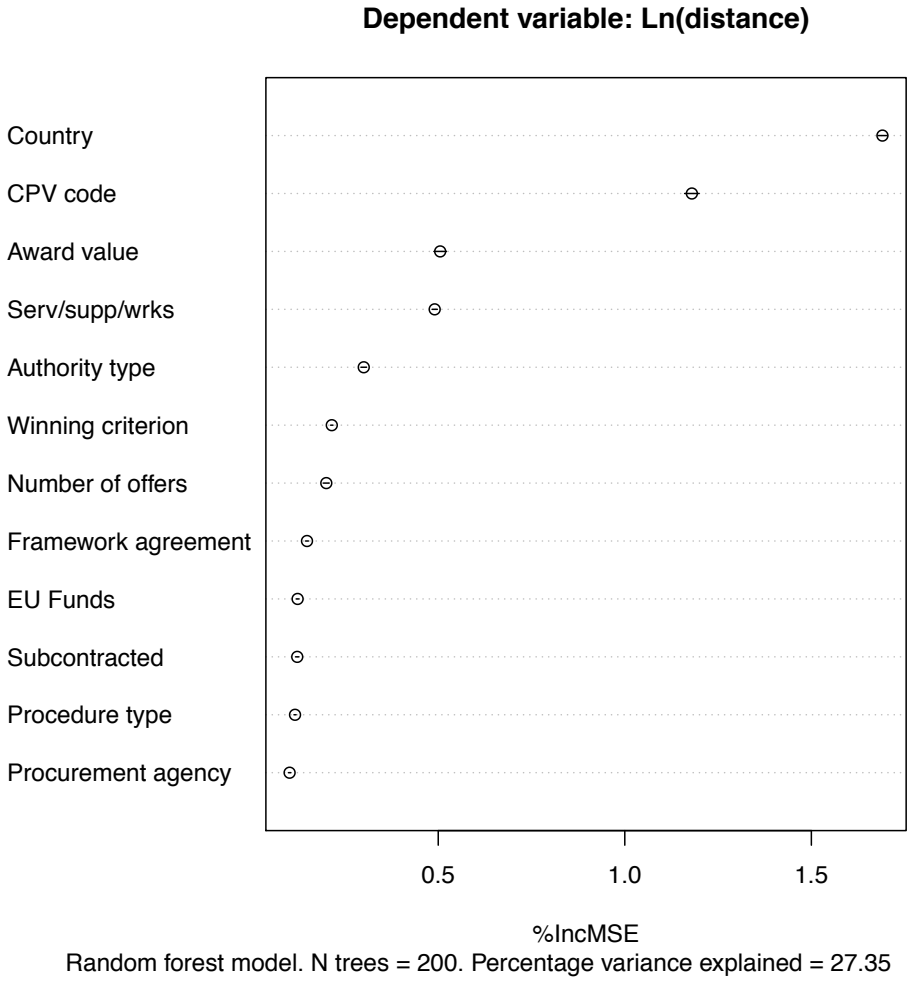


Figure 1: Variable importance plot

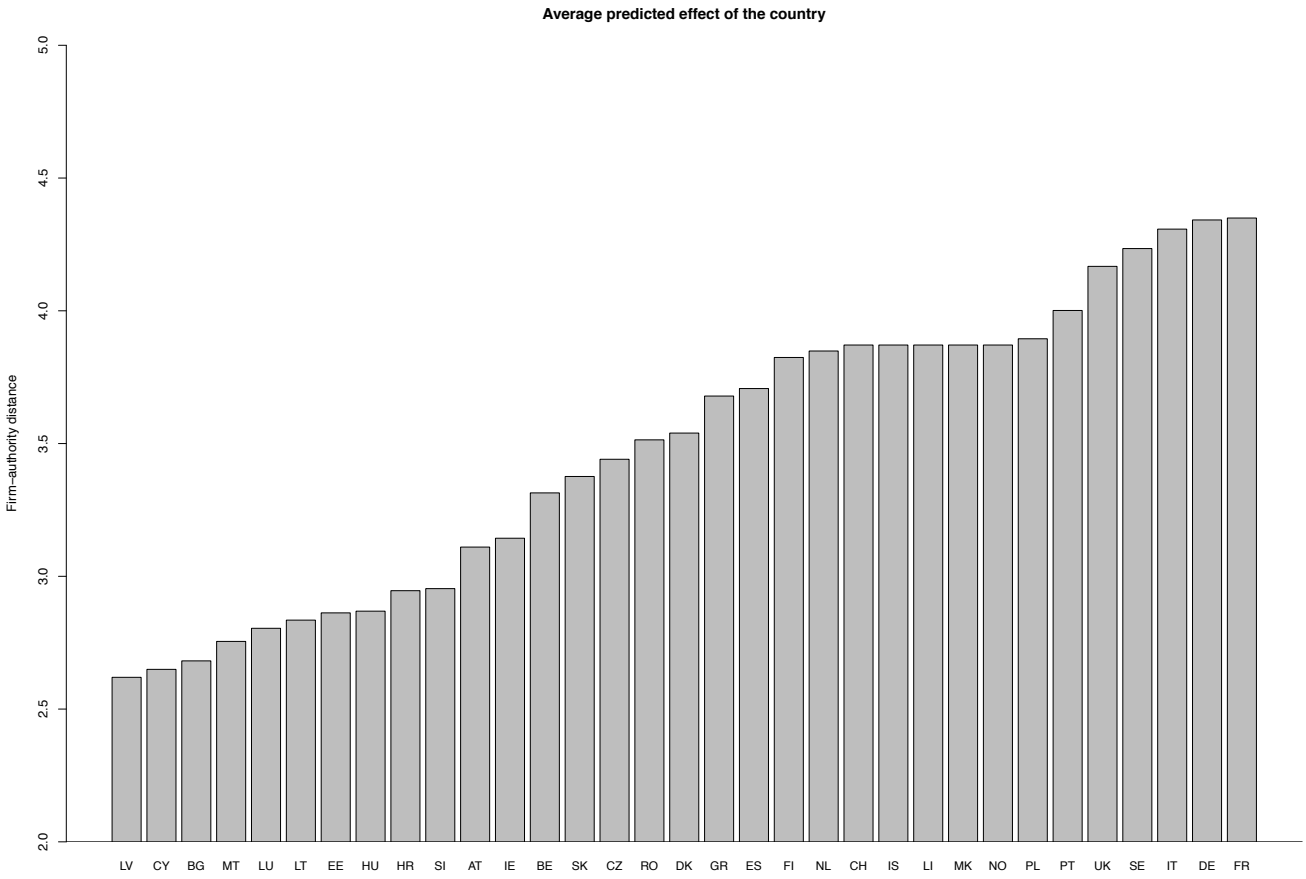


Figure 2: Predictive effect of the country

Appendix 6. Linear regression results

Table 6. 1 Linear regression model with firm-authority matches dependent variable.

Variable	Coefficient	P-value
CPV code		
Telecommunications services	-0.4379	.00
Secondary education services	-.04194	.00
Postal and telecommunications services	-.399	.00
Installation of medical equipment	-.3982	.00
Road transport services	.0613	.00
Forestry services	.1063	.00
Reinsurance services	.1633	.01
Professional services for oil industry	.1700	.03
Nature of the product		
Services (baseline)		
Supplies	-0.002	.43
Works	-0.024	.00
Type of authority		
National govt (baseline)		
Local authorities	-0.057	.00
Utilities	0.020	.00
EU institutions	-0.014	.04
International organizations	0.081	.00
Body governed by public law	-0.021	.00
Other	-0.024	.00
National agency	0.007	.05
Local agency	-0.007	.03
Not specified	-0.033	.00
Size of contract award		
871000+	-.014	.00
<871000	-.007	.00
<322000 (baseline)		
<158000	.031	.00
Missing	.040	.00
<76300	.094	.00
<36200	.085	.00
<16600	.098	.00
<6940	.109	.00
<2270	.120	.00

<404		.273	.00
Framework agreement			
No (baseline)			
Yes	.037		.00
Subcontracting likely			
Missing (baseline)			
No		.012	.09
Yes		-.027	.00
Procurement agency			
Missing (baseline)			
No		0.006	.00
Yes		0.044	.00
Country			
FR		-0.201	.00
SE		-0.170	.00
UK		-0.151	.00
SI		-0.138	.00
NL		-.126	.00
NO		0.124	.00
FI		-.114	.00
DE		-.083	.00
ES		-.072	.00
IS		-.068	.00
DK		-.057	.00
IT		-.030	.00
IE		-.029	.00
PL		-.022	.00
GR		-.017	.00
BE		.006	.32
RO		.010	.07
CZ		.023	.00
BG		.028	.00
CH		.037	.01
EE		.037	.00
LU		.058	.00
PT		.060	.00
HU		.066	.00
LI		.073	.10
SK		.081	.00
MT		.086	.00
LT		.087	.00
MK		.115	.00

CY	.130	.00
HR	.141	.00
LV	.184	.00
Procedure type		
Accelerated negotiated (baseline)		
Accelerated restricted	-0.036	.14
Awarded without publication	-0.046	.00
Competitive dialogue	-0.087	.00
Missing	-0.070	.00
Negotiated with call	-0.023	.01
Negotiated without call	0.083	.00
Open	-0.026	.00
Restricted	-0.039	.00
The number of offers		
1	.008	.00
2 (baseline)		
3-4	-.015	.00
5-7	-.033	.00
8+	-.074	.00
MISS	-.032	.00
EU funding		
Missing(baseline)		
No	.002	.09
Yes	-.027	.00
Criterion for deciding winner		
Lowest price (baseline)		
Most economical offer	-.010	.00
Missing	-.008	.00
Firm contract count	.114	.00
Authority contract count	.077	.00
Intercept	-.047	.03
N	1467677	

Table 6.2. Linear regression results on distance data.

Variable	Coefficient	P-value
CPV code		
Cybercafe services	-3.014	.13
Primary education services	-2.302	.00
Recreational, cultural, services	-1.130	.00
Real estate services	-1.116	.00
Pipeline inspection services	.857	.00
Apiculture services	1.250	.11
Leather	1.411	.03
Trailers and semi-trailers for agriculture	1.454	.14
Nature of the product		
Services (baseline)		
Supplies	.385	.00
Works	.058	.00
Type of authority		
National govt (baseline)		
Local authorities	0.079	.00
Utilities	0.415	.00
EU institutions	0.554	.00
International organizations	0.397	.00
Body governed by public law	0.123	.00
Other	0.175	.00
National agency	0.037	.01
Local agency	0.145	.00
Not specified	0.193	.00
Size of contract award		
871000+	-.024	.00
<871000	-.025	.00
<32200 (baseline)		
<158000	.004	.55
<76300	.043	.00
Missing	-.012	.06
<36200	.096	.00
<16600	.014	.00
<6940	.185	.00
<2270	.273	.00
<404	.112	.00

Framework agreement		
No (baseline)		
Yes	-0.099	.00
Subcontracting likely		
Missing (baseline)		
No	.025	.00
Yes	.023	.00
Procurement agency		
Missing (baseline)		
No	-.032	.00
Yes	.085	.00
Country		
CY	-1.806	.00
MT	-1.651	.00
BG	-0.961	.00
LV	-0.913	.00
LT	-0.741	.00
HU	-0.645	.00
HR	-0.629	.00
LU	-0.623	.00
EE	-0.473	.00
SI	-0.343	.00
IE	-0.102	.00
RO	-0.090	.00
SK	-0.023	.40
BE	0.036	.08
CZ	0.129	.00
GR	0.327	.00
ES	0.344	.00
PL	0.426	.00
DK	0.439	.00
FI	0.494	.00
PT	0.569	.00
NL	0.763	.00
UK	0.966	.00
IT	0.987	.00
FR	1.080	.00
SE	1.164	.00
DE	1.271	.00

Procedure type		
Accelerated negotiated (baseline)		
Accelerated restricted	-.054	.17
Awarded without publication	.132	.00
Competitive dialogue	.410	.00
Missing	.398	.00
Negotiated with call	.080	.02
Negotiated without call	.130	.00
Open	.118	.00
Restricted	.105	.00
The number of offers		
1	-.093	.00
2 (baseline)		
3-4	.041	.00
5-7	.064	.00
8+	.084	.00
MISS	.038	.00
EU funding		
Missing(baseline)		
No	-.0101	.00
Yes	.2165	.00
Criterion for deciding winner		
Lowest price (baseline)		
Most economical offer	-.014	.00
Missing	.269	.00
Intercept	2.937	.00
N	1267239	